กระบวนการและแบบจำลองสำหรับการคัดกรองเพื่อการจัดการคุณภาพข้อมูลในคราวด์ซอร์สซิงแพลตฟอร์ม

นายกฤตย์ กังวาลพงศ์พันธุ์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2565

Processes and Screening Models for Data Quality Management in a Crowdsourcing Platform

Mr. Krit Gangwanpongpun

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Engineering in Computer Engineering
Department of Computer Engineering
FACULTY OF ENGINEERING
Chulalongkorn University
Academic Year 2022

หัวข้อวิทยานิพนธ์	กระบวนการและแบบจำลองสำหรับการคัดกรองเพื่อการจัดการคุณภาพข้อมูลในคราวด์ซอร์สซิงแพลตฟอร์ม
โดย	นายกฤตย์ กังวาลพงศ์พันธุ์
สาขาวิชา	วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	อาจารย์ ดร.เอกพล ช่วงสุวนิช
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม	รองศาสตราจารย์ ดร.โปรดปราน บุณยพุกกณะ

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต

---------- คณบดีคณะวิศวกรรมศาสตร์
(ศาสตราจารย์ ดร.สุพจน์ เตชวรสินสกุล)

คณะกรรมการสอบวิทยานิพนธ์

---------- ประธานกรรมการ
(รองศาสตราจารย์ ดร.อติวงศ์ สุชาโต)

---------- อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(อาจารย์ ดร.เอกพล ช่วงสุวนิช)

---------- อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม
(รองศาสตราจารย์ ดร.โปรดปราน บุณยพุกกณะ)

---------- กรรมการภายนอกมหาวิทยาลัย
(รองศาสตราจารย์ ดร.สรณะ นุชอนงค์)

กฤตย์ กังวาลพงศ์พันธุ์ : กระบวนการและแบบจำลองสำหรับการคัดกรองเพื่อการจัดการคุณภาพข้อมูลในคราวด์ซอร์สซิงแพลตฟอร์ม. (Processes and Screening Models for Data Quality Management in a Crowdsourcing Platform) อ.ที่ปรึกษาหลัก : อ. ดร.เอกพล ช่วงสุวนิช, อ.ที่ปรึกษาร่วม : รศ. ดร.โปรดปราน บุณยพุกกณะ

การเก็บรวบรวมข้อมูลด้วยคราวด์ซอร์สซิงเป็นวิธีที่โดยทั่วไปมีความเร็วมากกว่า มีต้นทุนต่ำกว่า และมีความหลากหลายมากกว่าวิธีการเก็บรวบรวมข้อมูลแบบอื่น ๆ อย่างไรก็ตาม คราวด์ซอร์สซิงอาจเผชิญกับปัญหาคุณภาพ เช่น การติดป้ายกำกับผิดหรือการนำมาใช้ในทางที่ไม่เหมาะสม ดังนั้น กระบวนการควบคุมคุณภาพเป็นสิ่งที่จำเป็นสำหรับแพลตฟอร์มคราวด์ซอร์สซิง วิทยานิพนธ์นี้ศึกษาค้นคว้าอุปสรรคและวิธีการแก้ไขที่เป็นไปได้ในการจัดการคุณภาพของผู้ใช้งานแพลตฟอร์มคราวด์ซอร์สซิง ส่วนแรกเน้นวิธีการเพิ่มกระบวนการในคราวด์ซอร์สซิง โดยศึกษา 3 วิธี ได้แก่ 1. งานที่จำเป็นต้องทำก่อน 2. คำถามมาตรฐานแบบทองคำ และ 3. การทำซ้ำของข้อมูล พบว่างานที่จำเป็นต้องทำก่อนเป็นสิ่งจำเป็นเพื่อคัดกรองให้ได้ผู้ปฏิบัติงานที่มีคุณภาพสูง โดยควรเน้นไปที่ลักษณะเฉพาะและรายละเอียดของงาน คำถามที่ตรวจสอบความสอดคล้องระหว่างงานดีกว่าคำถามแบบชัดเจนในการตรวจสอบด้วยคำถามมาตรฐานทองคำ ผู้ตรวจสอบข้อมูลคนเดียวอาจนำไปสู่การปรับปรุงคุณภาพข้อมูลได้มากที่สุด ส่วนที่สองคือ การใช้แบบจำลองการเรียนรู้ของเครื่องที่ใช้ข้อมูลพฤติกรรมในการทำนายคุณภาพของข้อมูล ซึ่งวิธีนี้ยังช่วยคัดกรองข้อมูลคุณภาพต่ำออกไปได้โดยไม่เสียทรัพยากรเพิ่มเติม

สาขาวิชา	วิศวกรรมคอมพิวเตอร์	ลายมือชื่อนิสิต .....
ปีการศึกษา	2565	ลายมือชื่อ อ.ที่ปรึกษาหลัก .....
		ลายมือชื่อ อ.ที่ปรึกษาร่วม .....

6070108321 : MAJOR COMPUTER ENGINEERING

KEYWORD: crowdsourcing, data labeling, data quality control

Krit Gangwanpongpun : Processes and Screening Models for Data Quality Management in a Crowdsourcing Platform. Advisor: Ekapol Chuangsuwanich, Ph.D. Co-advisor: Assoc. Prof. PROADPRAN PUNYABUKKANA, Ph.D.

Crowdsourcing is generally a faster, more cost-effective, and diverse method of data collection. However, crowdsourcing might suffer from quality issues such as mislabeling or abuse. Thus, a quality control process is necessary for any crowdsourcing platform. This thesis explores the challenges and possible solutions in user quality management for crowdsourcing platforms. The first part focuses on augmenting the crowdsourcing process. Three aspects were studied: 1. Job Prerequisites, 2. Gold Standard Questions, and 3. Data Redundancy. I have found that job prerequisites are necessary to screen for high-quality workers, and emphasis should be put on the task specifics. Questions that check for consistency between tasks are better than obvious questions as a gold standard question. A single data validator may yield most of the improvement in data quality. The second part is about using mchine learning models that utilize behavioral data to predict the quality of data. This approach can also help screen out low-quality data witout requiring additional resources.

Field of Study: Computer Engineering
Academic Year: 2022

Student's Signature .....
Advisor's Signature .....
Co-advisor's Signature .....

กิตติกรรมประกาศ

ขอกราบขอบพระคุณ อาจารย์ ดร.เอกพล ช่วงสุวนิช และ รองศาสตราจารย์ ดร.โปรดปราน บุญยพุกกณะ อาจารย์ที่ปรึกษา และอาจารย์ที่ปรึกษาร่วม ที่แนะนำแนวทาง และให้ความช่วยเหลือในการทำวิจัยจนทำให้วิทยานิพนธ์นี้สำเร็จลุล่วงไปได้ด้วยดี

ขอกราบขอบพระคุณ รองศาสตราจารย์ ดร.อติวงศ์ สุชาโต และ รองศาสตราจารย์ ดร.สรณะ นุชอนงค์ ที่ได้สละเวลามาเป็นคณะกรรมการสอบวิทยานิพนธ์ และได้กรุณาให้คำแนะนำต่าง ๆ เพื่อทำให้วิทยานิพนธ์นี้มีคุณภาพมากยิ่งขึ้น

ขอกราบขอบพระคุณ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ที่สนับสนุนทุนอุดหนุนการศึกษาระดับบัณฑิตศึกษาสำหรับนิสิตเก่าวิศวกรรมศาสตรบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย

และสุดท้ายขอกราบขอบพระคุณ คณาจารย์และเจ้าหน้าที่ภาควิชาวิศวกรรมคอมพิวเตอร์ทุกท่าน ทุกคนในห้องปฏิบัติการ Spoken Language Systems and Assistive Technology และห้องปฏิบัติการข้างเคียง บุคคลรอบตัว ครอบครัว บรรดาเพื่อน รุ่นพี่ รุ่นน้อง ไม่ว่าใกล้หรือไกล ที่ได้ให้ความรู้ ความช่วยเหลือ ชี้แนะแนวทาง ให้คำแนะนำ และให้กำลังใจตลอดระยะเวลาในการทำวิทยานิพนธ์นี้

กฤตย์ กังวาลพงศ์พันธุ์

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1 บทนำ

1.1. ความเป็นมาและความสำคัญ

การทำงานด้าน**การเรียนรู้ของเครื่อง (Machine Learning)** ต้องใช้ข้อมูลปริมาณมากใน**การสอนแบบจำลอง (train model)** ส่งผลให้การเก็บข้อมูลของนักวิจัยและผู้ที่ทำงานด้านการเรียนรู้ของเครื่องเป็นงานที่ต้องใช้ระยะเวลานานและงบประมาณที่สูง รวมถึงบางครั้งยังไม่สามารถหาข้อมูลได้ในปริมาณตามต้องการ แต่ด้วยระบบ**คราวด์ซอร์สซิง (Crowdsourcing)** ซึ่งเข้ามามีบทบาทในการช่วยเก็บข้อมูลในยุคนี้ ส่งผลให้นักวิจัยสามารถเข้าถึงกลุ่มคนจำนวนมากได้อย่างสะดวก เก็บข้อมูลในปริมาณที่ต้องการได้อย่างรวดเร็วด้วยงบประมาณที่น้อยลง โดย**คราวด์ซอร์สซิงแพลตฟอร์ม (Crowdsourcing Platform)** เป็นแพลตฟอร์มที่อาศัยแนวคิดการเข้าถึงกลุ่มผู้ใช้งานจำนวนมากบนอินเทอร์เน็ตเพื่อรวมกำลังทำสิ่งที่ต้องการให้สำเร็จ ซึ่งในปัจจุบันมีคราวด์ซอร์สซิงแพลตฟอร์มหลายแห่ง เช่น microWorkers, clickworker, figure-eight (crowdflower) และ mturk (Amazon Mechanical Turk) ซึ่งเป็นแพลตฟอร์มที่ผู้ทำงานด้านการเรียนรู้ของเครื่องใช้กันอย่างแพร่หลาย โดยแต่ละแพลตฟอร์มจะมี**บริการคราวด์ซอร์สซิง (Crowdsourcing Services)** ที่แตกต่างกันไป เช่น **งานขนาดเล็ก (Microworks/Microtasks)** ซึ่งเป็น**งานอย่างง่าย (simple tasks)** การรวบรวมเงินลงทุน **คราวด์ฟันดิง (Crowdfunding)** การแปลภาษา หรือบางแพลตฟอร์มก็มีหลายบริการให้เลือกใช้ [1]

ระบบคราวด์ซอร์สซิง เป็นระบบที่ให้ผู้ที่ต้องการข้อมูล เรียกว่า **ผู้ว่าจ้าง (Employers)** หรือ**ผู้ร้องขอ (Requesters)** สามารถใช้ในการเรียกเก็บข้อมูล สอบถามข้อมูล ร้องขอให้ช่วยจัดเรียง รวมถึงติดป้ายให้กับข้อมูล ผ่านการสร้าง**งาน (Task)** เข้าไปในคราวด์ซอร์สซิงแพลตฟอร์ม โดยเมื่อผู้ที่ต้องการข้อมูลทำการสร้างงานเข้ามาในระบบแล้ว ระบบจะทำการจัดแบ่งงานนั้นออกเป็นส่วนย่อย ๆ ที่เรียกว่า **งานขนาดเล็ก (Microworks/Microtasks)** เพื่อกระจายไปยัง**ผู้ปฏิบัติงาน (Workers)** ซึ่งเป็นผู้ใช้งานอินเทอร์เน็ตทั่วไปที่สมัครสมาชิกเข้าไปในระบบ เพื่อให้ผู้ปฏิบัติงานสามารถเข้ามาช่วยกันทำงานได้ และได้รับค่าตอบแทนกลับไปตามที่ผู้ที่ต้องการข้อมูลกำหนดไว้ให้แต่ละงานขนาดเล็ก

เนื่องจากการเก็บข้อมูลด้วยคราวด์ซอร์สซิงแพลตฟอร์ม เป็นการแบ่งงานจำนวนมากออกเป็นส่วนย่อย ๆ แล้วกระจายงานขนาดเล็กเพื่อเข้าถึงคนบนอินเทอร์เน็ตจำนวนมากให้ช่วยกันทำ ส่งผลให้ผู้ปฏิบัติงานสามารถเป็นใครก็ได้ คุณภาพของงานที่ได้จึงมีความหลากหลาย ตั้งแต่ต่ำกว่ามาตรฐานที่ผู้ร้องขอต้องการ จนถึงได้คุณภาพตามที่ผู้ร้องขอต้องการ เห็นได้ว่า คุณภาพของ**ข้อมูลผลลัพธ์ (data result)** ที่ได้จากการเก็บข้อมูลแบบคราวด์ซอร์สซิงไม่อาจรับประกันคุณภาพได้ทั้งหมด แต่คุณภาพของข้อมูลผลลัพธ์กลับเป็นปัจจัยสำคัญที่ส่งผลต่อผลงานวิจัยของนักวิจัยเป็นอย่าง

มาก ดังนั้นจึงมีผู้พยายามสร้างวิธีการตรวจสอบคุณภาพของข้อมูลผลลัพธ์และคัดกรองข้อมูลที่ไม่มีคุณภาพออกไป

การตรวจสอบคุณภาพและคัดกรองข้อมูลผลลัพธ์ที่ไม่มีคุณภาพที่ถูกสร้างขึ้นมามีหลากหลายรูปแบบ แบบดั้งเดิมที่สุดคือ**การทำซ้ำ (Redundant)** ของข้อมูล โดยให้งานขนาดเล็กหนึ่ง ๆ มีผู้ปฏิบัติงานมากกว่าหนึ่งคนทำ และใช้**วิธีนับเสียงข้างมาก (Majority vote)** ในการเลือกผลการปฏิบัติงานหรือคำตอบสุดท้าย หรือใช้**วิธีการสร้างมาตรฐานทองคำ (Gold Standard)** ขึ้นมาเพื่อตรวจสอบว่าคุณภาพของงานขนาดเล็กที่ผู้ปฏิบัติงานตอบมานั้น เป็นไปตามที่ต้องการหรือไม่ ซึ่งวิธีเหล่านี้ทั้งหมดล้วนส่งผลให้สูญเสียทรัพยากรไป ไม่ว่าจะเป็นการทำซ้ำของข้อมูลหรือการสร้างคำถามซึ่งเป็นมาตรฐานทองคำแทรกเข้าไป ส่งผลให้ผู้ร้องขอสูญเสียทรัพยากรกำลังคนของผู้ปฏิบัติงานไป และการทำงานขนาดเล็กแต่ละครั้งของผู้ปฏิบัติงานก็ต้องใช้เวลามากขึ้น แทนที่ในกำลังคนและเวลาที่มีอยู่จะได้ข้อมูลในปริมาณที่มากขึ้น กลับต้องสูญเสียไปเพื่อใช้ในการตรวจสอบคุณภาพของผลการปฏิบัติงาน นักวิจัยที่สร้างการตรวจสอบคุณภาพและคัดกรองงานจึงหันมาใช้กรรมวิธีการตรวจจับ**พฤติกรรม (behavior)** ของผู้ปฏิบัติงาน และสร้าง**แบบจำลอง (model)** พฤติกรรมเพื่อ**ทำนาย (predict)** และ**จำแนก (classification)** คุณภาพของข้อมูลผลลัพธ์จากการปฏิบัติงานนั้นแทนวิธีการแบบเก่า

ผู้วิจัยได้สร้างคราวด์ซอร์สซิงแพลตฟอร์มชื่อ “ว่าง” ขึ้นเพื่อให้นักวิจัยในประเทศไทยได้ใช้เป็นช่องทางในการเก็บข้อมูล ซึ่ง “ว่าง” มีเครื่องมือที่ช่วยให้นักวิจัยสามารถสร้างงานเข้าไปในแพลตฟอร์มได้อย่างสะดวก ซึ่งนอกจากการสร้างงานแล้ว “ว่าง” ยังมีเป้าหมายที่จะอำนวยความสะดวกให้แก่นักวิจัยในการตรวจสอบคุณภาพของข้อมูลผลลัพธ์ที่ได้กลับไปด้วย ดังนั้นผู้วิจัยจึงต้องการพัฒนาวิธีการตรวจสอบคุณภาพและคัดกรองงาน แต่งานวิจัยที่พัฒนาการตรวจสอบคุณภาพและคัดกรองงานด้วยการตรวจจับพฤติกรรมที่มีอยู่ในปัจจุบัน ต่างศึกษาและทดลองการสร้างแบบจำลองแล้วนำมาตรวจสอบคุณภาพเพียงแค่งานหนึ่ง ๆ เท่านั้น ยังไม่มีการศึกษาทดลอง เพื่อทดสอบและพัฒนาให้สามารถใช้กระบวนการฝึกฝนแบบจำลองเดิมในการตรวจสอบงานที่หลากหลายได้ แต่ “ว่าง” เป็นคราวด์ซอร์สซิงแพลตฟอร์มที่รองรับการสร้างงานได้หลายรูปแบบ จึงจำเป็นต้องสร้างกระบวนการฝึกฝนแบบจำลองที่สามารถประยุกต์ใช้กับงานได้หลายชนิด จึงพิจารณาทดลองกระบวนการสร้างแบบจำลองจากพฤติกรรมเพื่อนำมาใช้ตรวจสอบคุณภาพในงานหลากหลายรูปแบบ และศึกษาผลลัพธ์ของการใช้งานจริงในสถานการณ์จำลองที่แตกต่างกัน เพื่อดูว่า การใช้งานแบบใดให้ประสิทธิภาพในการคัดกรองที่เป็นที่พึงพอใจของผู้ร้องขอข้อมูล และเสียทรัพยากรมในการเก็บข้อมูลและตรวจสอบคุณภาพของข้อมูลน้อยที่สุด

เอกสารฉบับนี้มีโครงสร้างดังต่อไปนี้ ส่วนทฤษฎีที่เกี่ยวข้อง อธิบายทฤษฎีที่ใช้ประกอบในงานวิจัย ส่วนงานวิจัยที่เกี่ยวข้อง บอกเกี่ยวกับงานวิจัยที่ใกล้เคียงกัน ซึ่งอธิบายถึงงานที่นักวิจัยอื่น

ศึกษามาก่อนในขอบข่ายงานเดียวกัน รวมถึงบ่งบอกถึงความสำคัญของปัญหา ส่วนวัตถุประสงค์ของการวิจัย บอกเป้าหมายของงานวิจัยฉบับนี้ ส่วนขอบเขตการวิจัย บ่งบอกถึงข้อจำกัดและขอบเขตที่งานวิจัยฉบับนี้จะศึกษา จากนั้นบอกถึงขั้นตอนการดำเนินงานวิจัยไว้ในส่วนของขั้นตอนการวิจัย และบอกถึงประโยชน์ที่ได้รับจากงานวิจัยฉบับนี้ ไว้ในส่วนของประโยชน์ที่คาดว่าจะได้รับ ส่วนแนวคิดและวิธีการวิจัย อธิบายถึงวิธีการที่งานวิจัยฉบับนี้เสนอ ทดลอง และใช้ในการแก้ปัญหา ต่อมาเป็นส่วนสรุปผลการวิจัยและข้อเสนอแนะ เพื่อสรุปและอภิปรายผลการวิจัย และเสนอแนะแนวทางการศึกษาต่อในอนาคต สุดท้ายเป็นแหล่งอ้างอิงถึงงานวิจัย เอกสารและหนังสือฉบับอื่น ในส่วนบรรณานุกรม

1.2. วัตถุประสงค์การวิจัย

เพื่อศึกษากระบวนการตรวจสอบคุณภาพของข้อมูลด้วยการเพิ่มกระบวนการ และกระบวนการสร้างแบบจำลองที่ใช้ได้ทั่วไปสำหรับการตรวจหาคำตอบคุณภาพต่ำสำหรับงานหลายประเภทในระบบคราวด์ซอร์สซิง โดยใช้พฤติกรรมในการทำงานของผู้ปฏิบัติงานในการตรวจสอบ

1.3. ขั้นตอนการวิจัย

1. ศึกษาการทำงานของระบบคราวด์ซอร์สซิง และพัฒนาระบบคราวด์ซอร์สซิง “ว่าง”
2. ศึกษาทฤษฎีที่เกี่ยวข้อง
3. ศึกษาค้นคว้างานวิจัยที่เกี่ยวข้อง
4. ออกแบบแนวคิดและวิธีการวิจัย
5. จัดเตรียมงานและข้อมูลผลลัพธ์เพื่อใช้เก็บข้อมูลพฤติกรรมการทำงาน
6. เก็บข้อมูลพฤติกรรมการทำงานของผู้ปฏิบัติงานด้วยงานที่เตรียมไว้
7. นำข้อมูลพฤติกรรมที่เก็บได้มาทำการทดลองในสถานการณ์จำลองต่าง ๆ ที่กำหนด รวมถึงฝึกฝนแบบจำลองตามแนวคิดและวิธีการวิจัยที่เตรียมไว้
8. ตรวจสอบผลลัพธ์ ความแม่นยำของแบบจำลอง และประโยชน์ของการนำไปประยุกต์ใช้ในสถานการณ์จริงต่าง ๆ หากยังไม่เป็นที่น่าพอใจ ให้ปรับเปลี่ยนรายละเอียดการทดลองในขั้นตอนต่าง ๆ เพื่อปรับปรุงให้ได้ผลลัพธ์ที่สามารถอภิปรายและนำไปประยุกต์ใช้จริงได้ เช่น
    1. ปรับเปลี่ยนจำนวนมิติ จำนวนชั้นของแบบจำลอง
    2. ปรับเปลี่ยนค่าขีดแบ่งที่ใช้ตัดสินใจแบ่งกลุ่มพฤติกรรมในช่วงการฝึกฝนแบบจำลอง
9. ทดสอบและวัดผลความแม่นยำของแบบจำลองสุดท้าย
10. สรุปผลการวิจัย
11. เผยแพร่งานวิจัย
12. จัดทำเอกสารวิทยานิพนธ์

1.4. ขอบเขตการวิจัย

1. แบบจำลองในงานวิจัยนี้ใช้ตรวจสอบเฉพาะงานที่บุคคลทั่วไปสามารถทำได้โดยไม่ต้องใช้ความรู้เฉพาะทางหรือทักษะพิเศษอื่นใดในการทำงาน เช่น การระบุชนิดวัตถุในภาพ เมื่อวัตถุในภาพเป็นสิ่งของทั่วไป การถอดความเสียงเป็นข้อความในกรณีที่ผู้ปฏิบัติงานเป็นเจ้าของภาษาและเนื้อหาไม่ได้มีศัพท์เฉพาะทาง ดังนั้นจึงเป็นงานที่หากผู้ปฏิบัติงานทำงานอย่างมีคุณภาพ ข้อมูลผลลัพธ์ที่ได้ต้องเป็นคำตอบที่ถูกต้อง
2. แบบจำลองครอบคลุมเฉพาะการตรวจสอบคุณภาพของข้อมูลผลลัพธ์ โดยนับว่าข้อมูลผลลัพธ์นั้นมีคุณภาพสูงก็ต่อเมื่อข้อมูลผลลัพธ์ที่ได้คือคำตอบที่ถูกต้องของงานนั้น ๆ ดังนั้นไม่ว่าผู้ปฏิบัติงานจะมั่วคำตอบ โกง ตอบคำตอบเดียวตลอดงาน หรือการกระทำอื่นใดที่ส่งผลให้คำตอบผิด จะถือว่าเป็นข้อมูลผลลัพธ์ที่ไม่มีคุณภาพ
3. แบบจำลองนี้จะครอบคลุมในส่วนงานอย่างง่ายที่ไม่ซับซ้อนหรือไม่ต้องใช้ความคิดสร้างสรรค์สูงในการทำงาน เช่น การระบุชนิดวัตถุในรูป การถอดความเสียงจากไฟล์เสียงหรือจากวิดีทัศน์เป็นข้อความ การตอบคำถามตามคำสั่ง
4. แบบจำลองนี้ใช้ข้อมูลพฤติกรรมและเหตุการณ์ที่สามารถเก็บได้จากเว็บเบราว์เซอร์เท่านั้น
5. แบบจำลองนี้จะสามารถตรวจจับข้อมูลผลลัพธ์คุณภาพต่ำได้ในงานที่มีตัวอย่างพฤติกรรมในการทำงานเป็นต้นแบบในการเปรียบเทียบเท่านั้น

1.5. ประโยชน์ที่ได้รับ

1. ได้แนวทางการประยุกต์ใช้ ข้อดี การแลกเปลี่ยน สิ่งที่ควรคำนึงถึง ของการเพิ่มกระบวนการและการใช้แบบจำลองการเรียนรู้ของเครื่องเข้ามาใช้ตรวจสอบข้อมูลคุณภาพต่ำของการเก็บข้อมูลด้วยคราวด์ซอร์สซิง
2. ได้กระบวนการฝึกฝนแบบจำลองที่สามารถใช้ได้ทั่วไปในการตรวจหาคำตอบคุณภาพต่ำสำหรับงานหลายประเภทในระบบคราวด์ซอร์สซิง โดยใช้พฤติกรรมในการทำงานของผู้ปฏิบัติงานในการตรวจสอบ
3. แบบจำลองสามารถช่วยลดภาระงานในการตรวจสอบข้อมูลผลลัพธ์และคำตอบที่ได้จากระบบคราวด์ซอร์สซิง รวมถึงช่วยให้ได้ข้อมูลผลลัพธ์ที่มีคุณภาพสูงขึ้นในค่าใช้จ่ายที่ลดลง ส่งผลให้งานด้านการเรียนรู้ของเครื่องอื่นได้ข้อมูลผลลัพธ์ที่มีคุณภาพไปใช้ในงานต่อไป

1.6. ผลงานวิจัยที่ได้รับการเผยแพร่

Gangwanpongpun, K. Punyabukkana, P. and Chuangsuwanich E, *Analysis of the Trade-Off in Data Quality Management for Crowdsourcing*. In *2023 The 20th International Joint Conference on Computer Science and Software Engineering (JCSSE)*. 2023 IEEE

บทที่ 2 ทฤษฎีที่เกี่ยวข้อง

2.1. การถดถอยเชิงเส้น (Linear Regression)

การถดถอยเชิงเส้น เป็นการสร้างแบบจำลองทางสถิติชนิดหนึ่ง โดยมุ่งเป้าให้แบบจำลองที่เกิดขึ้นสามารถทำนายผลลัพธ์บางอย่างที่เราต้องการทราบจากข้อมูลที่เรามีอยู่ได้ โดยปกตินั้นการสร้างแบบจำลองสามารถใช้สมการได้หลายรูปแบบ แต่หากกล่าวถึงการถดถอยเชิงเส้นนั้นจะใช้สมการเชิงเส้นหลายตัวแปรเป็นแบบจำลอง โดยมีรูปแบบสมการพื้นฐานที่สุดดังนี้ [2]

$$E[Y|X_1 = x_1, X_2 = x_2, \dots, X_K = x_K] = \emptyset(x_1, x_2, \dots, x_K)$$

$$\emptyset(x_1, x_2, \dots, x_K) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_K$$

Y = ตัวแปรตาม (Dependent Variable) หรือตัวแปรผลลัพธ์ (Response Variable)

β_0 = จุดตัดแกน Y (Y intercept) เป็นพารามิเตอร์ (parameter) ของสมการ

$\beta_{1\dots K}$ = ประชากรสัมประสิทธิ์ความชัน (Population Slop Coefficient) เป็นพารามิเตอร์ (parameter) ของสมการ

$x_{1\dots K}$ = ตัวแปรต้น (Independent Variable/Explanatory Variable)

ค่า $Y, \beta_0, \beta_{1\dots K},$ และ $x_{1\dots K}$ สามารถเป็นได้ทั้งปริมาณสเกลาร์ (scalar) ซึ่งเป็นเลขเพียงตัวเดียว หรือเป็นเมทริกซ์ (matrix) ซึ่งประกอบจากเลขหลายตัวก็ได้ โดยจำต้องมีลักษณะของจำนวนตัวเลขและมิติของเมทริกซ์ที่สัมพันธ์กันทั้งสมการ

ผลลัพธ์ที่เราต้องการทำนาย ในที่นี้คือค่า Y หรือผลลัพธ์จากฟังก์ชัน $\emptyset$ เรียกว่า ตัวแปรตาม จะมีค่าเปลี่ยนไปตามข้อมูลที่เราทราบซึ่งใส่เข้าไปในฟังก์ชันเป็นตัวแปรต้นหรือตัวแปรอิสระ โดยมีปัจจัยสำคัญคือ ค่าพารามิเตอร์ ในแบบจำลอง ซึ่งเกิดจากการสอนแบบจำลองด้วยวิธีการถดถอยเชิงเส้น เพื่อให้แบบจำลองสามารถทำนายคำตอบในสิ่งที่เราต้องการออกมาได้ พารามิเตอร์ ถือเป็นสิ่งสำคัญของแบบจำลองที่จะส่งผลต่อความแม่นยำของผลลัพธ์ที่ได้จากแบบจำลอง ดังนั้นสิ่งที่สำคัญที่สุดในการสร้างแบบจำลองคือ การสอนแบบจำลอง เพื่อปรับค่าพารามิเตอร์ให้มีค่าที่ควรจะเป็น

การสร้างแบบจำลองด้วยการถดถอยเชิงเส้น ตอนเริ่มต้นเราต้องมีข้อมูลที่ครบคู่ระหว่างข้อมูลที่เป็นตัวแปรต้นและข้อมูลที่เป็นตัวแปรตาม โดยเราจะแบ่งข้อมูลนี้ออกเป็นสามส่วน ได้แก่

1.**ข้อมูลสำหรับการฝึกฝน (Training Data)** ใช้ในการปรับน้ำหนักพารามิเตอร์เพื่อสร้างสมการแบบจำลอง

2. **ข้อมูลสำหรับการตรวจสอบ (Validation Data)** ใช้ในการตรวจสอบว่าควรจะหยุดปรับน้ำหนักพารามิเตอร์หรือยัง ในช่วงการสร้างสมการแบบจำลอง

3. **ข้อมูลสำหรับการทดสอบ (Test Data)** ใช้ในการทดสอบประสิทธิภาพสุดท้ายของสมการแบบจำลอง หลังจากปรับน้ำหนักพารามิเตอร์และสร้างสมการแบบจำลองเสร็จ

โดยเราจะเริ่มจากการสร้างสมการที่ค่าพารามิเตอร์ถูกกำหนดขึ้นโดยการสุ่ม หลังจากนั้นใส่ค่าตัวแปรต้นจากกลุ่มข้อมูลสำหรับการฝึกฝนลงไป จะได้ผลลัพธ์ที่สมการทำนายออกมา ให้นำผลลัพธ์ที่ทำนายมาเทียบกับผลลัพธ์จริงที่เป็นตัวแปรตามที่ทราบอยู่แล้ว ด้วย**ฟังก์ชันสูญเสีย (Loss Function)** ที่กำหนดขึ้น ซึ่งเป็น**ฟังก์ชันระยะทาง (Distance Function)** ชนิดหนึ่ง เมื่อยังมีความหากกันอยู่ ต้องทำการปรับค่าน้ำหนักของพารามิเตอร์ ด้วย**กระบวนการสืบสายตามความลาดชัน (Gradient Descent)** ดังสมการ

$$\beta_{0_{new}} = \beta_{0_{old}} - \propto \frac{1}{n}\sum_{i=1}^{n}(\emptyset(x_i) - y_i)$$

$$\beta_{1_{new}} = \beta_{1_{old}} - \propto \frac{1}{n}\sum_{i=1}^{n}(\emptyset(x_i) - y_i) \cdot x_i$$

$\propto$ = อัตราการเรียนรู้ (learning rate)

n = จำนวนข้อมูลสำหรับการฝึกฝน

$x_{i\ldots n}$ = ข้อมูลที่เป็นตัวแปรต้นในกลุ่มข้อมูลสำหรับการฝึกฝน

$y_{i\ldots n}$ = ข้อมูลที่เป็นตัวแปรตามในกลุ่มข้อมูลสำหรับการฝึกฝน

การปรับน้ำหนักพารามิเตอร์ด้วยกระบวนการสืบสายตามความลาดชัน จะถูกทำวนซ้ำไปเรื่อย ๆ สังเกตได้ว่า เมื่อเรานำค่าตัวแปรต้นจากกลุ่มข้อมูลสำหรับตรวจสอบใส่ลงไปในสมการแล้วผลลัพธ์ที่ทำนายออกมาทำให้เกิดค่าความสูญเสียจากฟังก์ชันสูญเสียน้อยลงในทุกครั้งที่ทำการปรับน้ำหนักพารามิเตอร์ หากปรับแล้วส่งผลให้เกิดค่าความสูญเสียจากฟังก์ชันสูญเสียมากขึ้น ให้หยุดทำการปรับน้ำหนักพารามิเตอร์ แล้วย้อนกลับไปใช้ค่าพารามิเตอร์ที่ทำให้เกิดค่าความสูญเสียน้อยที่สุดถือเป็นอันจบกระบวนการฝึกฝนเพื่อสร้างสมการแบบจำลอง

เราสามารถวัดประสิทธิภาพของสมการแบบจำลองได้โดยนำค่าตัวแปรต้นจากกลุ่มข้อมูลสำหรับการทดสอบใส่ลงในสมการแบบจำลองที่ได้ แล้ววัดผลที่ทำนายออกมาจากสมการแบบจำลอง

กับผลตัวแปรตามจริงที่ทราบค่าอยู่แล้ว ก็จะสามารถทราบประสิทธิภาพในการทำนายของสมการแบบจำลองได้

2.2. ฟังก์ชันระยะทาง (Distance Function)

เป็นสมการทางคณิตศาสตร์สำหรับการหาระยะทางระหว่างจุดสองจุดบน**ปริภูมิ (Space)** เพื่อที่จะสามารถนำระยะทางนั้นมาเปรียบเทียบกันได้ ว่าคู่จุดใดมีระยะทางที่ห่างมากกว่ากัน โดยฟังก์ชันระยะทางต้องมีคุณสมบัติคือ ค่าระยะทางหรือความห่างที่คำนวณได้ต้องไม่ติดลบ การคำนวณระยะทางไปกลับต้องได้ค่าเท่ากัน และถ้าตำแหน่งเดียวกัน ระยะห่างที่คำนวณได้ต้องเป็นศูนย์

ฟังก์ชันระยะทางสามารถนำสมการทางคณิตศาสตร์มาประยุกต์ใช้ได้หลายแบบ เช่น

สมการระยะทางของยุคลิด (Euclidean distance)

$$d(p,q) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

สมการความเหมือนกันแบบโคโซน์ (Cosine similarity)

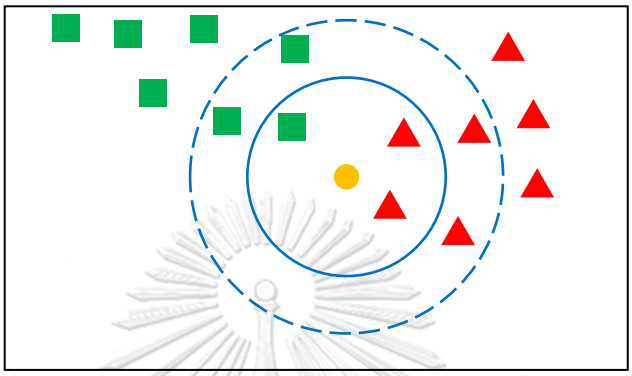
$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

2.3. การจำแนกประเภท (Classification)

เป็นกระบวนการในการจำแนกประเภทของข้อมูล ว่าข้อมูลนี้ควรถูกจัดจำแนกไปอยู่ในข้อมูลกลุ่มใด โดยการจำแนะประเภทมี**ระเบียบวิธี (Methodology)** ในการทำหลายวิธี เช่น ขั้นตอน**วิธีการเพื่อนบ้านที่ใกล้ที่สุด (K-Nearest Neighbors Algorithm: KNN) ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) การถดถอยโลจิสติกส์ (Logistic Regression)** หรือแม้แต่**การถดถอยเชิงเส้น (Linear Regression)** ก็สามารถใช้ในการจำแนกประเภทได้เช่นกัน หากสมการแบบจำลองสร้างมาเพื่อใช้ในการจำแนกประเภทของข้อมูล หรือแม้แต่การกำหนด**ค่าขีดแบ่ง (Threshold)** บางค่าขึ้นมา ก็สามารถใช้ในการจำแนกประเภทของข้อมูลได้เช่นเดียวกัน เช่น ถ้าค่าสายตา ติดลบเกิน 150 เซนติเมตร ถือว่าเป็นคนสายตาสั้น

ขั้นตอนวิธีการเพื่อนบ้านที่ใกล้ที่สุด (K-Nearest Neighbors Algorithm: KNN)

KNN เป็นวิธีการจำแนกประเภทของข้อมูลโดยพิจารณาข้อมูลเพื่อนบ้านที่อยู่ใกล้ที่สุด K ตัว โดยดูความใกล้จากการคำนวณระยะห่างด้วยฟังก์ชันระยะทาง แล้วตรวจสอบว่า เพื่อนบ้านทั้ง K ตัวนั้นอยู่ถูกจำแนกอยู่ในข้อมูลกลุ่มใดหรือ**คลาส (Class)** ใดบ้าง ทำการนับผลสำรวจ และสรุปว่าข้อมูลนี้จะถูกจำแนกให้อยู่คลาสเดียวกับคลาสที่มีจำนวนเพื่อนบ้านอยู่ในคลาสนั้นมากที่สุด



รูปภาพที่ 1 ภาพแสดงการจำแนกประเภทด้วยวิธีการเพื่อนบ้านที่ใกล้ที่สุด

2.4. การเลือกขีดแบ่ง (Threshold Selection)

ในบางครั้งการจำแนกประเภทของข้อมูลจำเป็นต้องอาศัยการกำหนดค่าขีดแบ่ง เพื่อใช้เป็นเกณฑ์ในการตัดสินว่า ข้อมูลนี้ควรถูกจัดอยู่ในประเภทใด หรือใช้ในการตัดสินว่าข้อมูลสองจุดนี้เป็นข้อมูลประเภทเดียวกันหรือไม่ เช่น เราสามารถใช้ค่าขีดแบ่งในการกำหนดว่า ข้อมูลสองจุดบนปริภูมิที่มีระยะห่างจากฟังก์ชันระยะทางเกินค่าขีดแบ่ง ถือว่าเป็นข้อมูลที่แตกต่างกัน แต่ถ้าไม่เกินค่าขีดแบ่งถือว่าเป็นข้อมูลที่มีความคล้ายคลึงกัน หรือสามารถถูกจัดจำแนกอยู่ในประเภทเดียวกันได้

การเลือกค่าขีดแบ่งเป็นสิ่งที่ต้องพิจารณาอย่างถี่ถ้วน เนื่องจากค่าขีดแบ่งที่เปลี่ยนไป จะส่งผลกระทบต่อความแม่นยำในการจำแนกประเภทของข้อมูลด้วย ดังนั้นเราจำเป็นต้องมีขั้นตอนการเลือกขีดแบ่งที่เหมาะสมกับงานของเรา ซึ่งการเลือกขีดแบ่งสามารถทำได้หลายวิธี หนึ่งในวิธีที่คนนิยมใช้ในการเลือกค่าขีดแบ่งสำหรับงานที่ต้องการแบ่งข้อมูลออกเป็นเพียงสองกลุ่ม หรือ **การจำประเภทแบบทวิภาค (Binary Classification)** คือ **เส้นโค้งรีซีฟเวอร์โอเปอเรติงแคแรกเทอริสติก (Receiver Operating Characteristic Curve: ROC curve)**

2.4.1. เส้นโค้งรีซีฟเวอร์โอเปอเรติงแคแรกเทอริสติก (Receiver Operating Characteristic Curve: ROC curve)

เป็นการ**วาด (plot)** เส้น**กราฟ (graph)** เพื่อแสดงผลความสามารถในการจำแนกประเภทเมื่อกำหนดค่าขีดแบ่งเป็นค่าต่าง ๆ ใช้ในการประกอบการตัดสินใจเพื่อเลือกค่าขีดแบ่งที่เหมาะสมกับงานนั้น โดยการวาดกราฟ จะวาดจาก**ค่าอัตราจริงบวก (true positive rate: TPR)** และ**อัตราเท็จ**

บวก (FPR) เกิดจากการนับผลของกรณีต่าง ๆ ที่จะเกิดขึ้น ซึ่งมีกรณีที่จะเกิดขึ้นพื้นฐานดังตารางต่อไปนี้

		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	True positive	False positive Type I error
	Predicted condition negative	False negative Type II error	True negative

ตารางที่ 1 ตารางแสดงค่าเมทริกซ์สับสน (Confusion Matrix)

จริงบวก (True positive) คือ ของจริงเป็นบวก ที่ทำนายเป็นบวก จึงทำนายถูก
เท็จลบ (False negative) คือ ของจริงเป็นบวก ที่ทำนายเป็นลบ จึงทำนายผิด
เท็จบวก (False positive) คือ ของจริงเป็นลบ ที่ทำนายเป็นบวก จึงทำนายผิด
จริงลบ (True negative) คือ ของจริงเป็นลบ ที่ทำนายเป็นลบ จึงทำนายถูก

โดยค่าที่ต้องใช้ในการวาดกราฟ มีสมการดังนี้

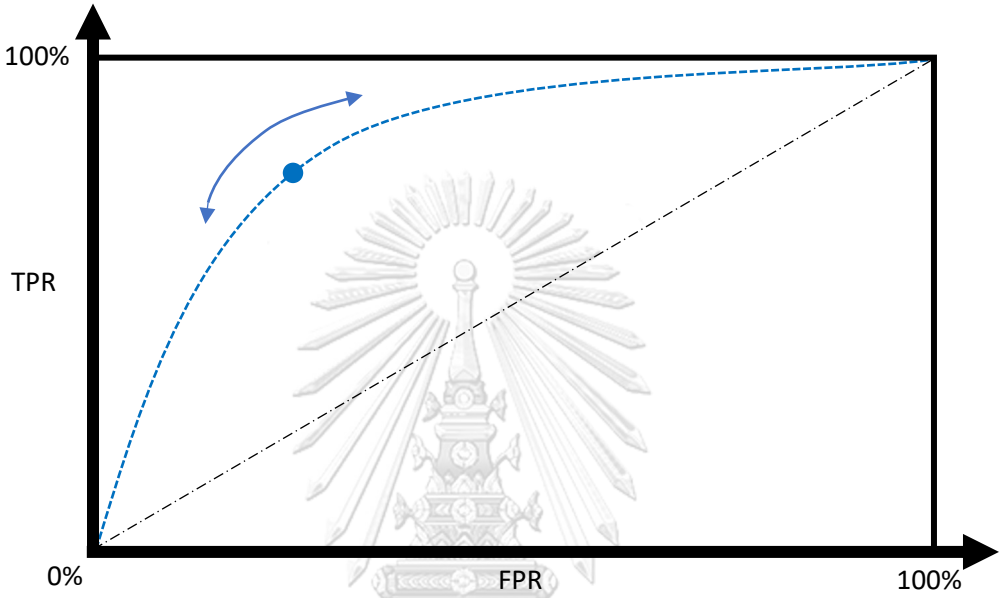
ค่าอัตราจริงบวก (true positive rate: TPR) หรือ ความไว (Sensitivity), รีคอล (recall) หรืออัตราการโดน (hit rate) คำนวณได้จาก

$$TPR = \frac{\sum True\ positive}{\sum Condition\ positive} = \frac{TP}{P} = \frac{TP}{TP + FN} = 1 - FNR$$

ค่าอัตราเท็จบวก (false positive rate: FPR) หรือ ค่าตกออก (Fall-out) คำนวณได้จาก

$$FPR = \frac{\sum False\ positive}{\sum Condition\ negative} = \frac{FP}{N} = \frac{FP}{FP + TN} = 1 - TNR$$

การวาดกราฟซึ่งเป็นคู่อันดับของค่าอัตราจริงบวกและค่าอัตราเท็จบวก ทำได้โดยเลือกค่าขีดแบ่งที่ตัวเลขต่าง ๆ แล้วใช้ค่าขีดแบ่งนั้นในการจำแนกประเภทข้อมูล เมื่อจำแนกเสร็จให้ทดสอบผลว่าการจำแนกที่ทำนายออกมาตรงกับความเป็นจริงหรือไม่ เข้ากรณีใดในกรณีพื้นฐานข้างต้น นับจำนวนที่เกิดขึ้นของทั้ง 4 กรณี นำมาคำนวณค่าอัตราจริงบวกและค่าอัตราเท็จบวกเพื่อจดลงไปบนกราฟ หลังจากนั้นเปลี่ยนค่าขีดแบ่งเป็นค่าอื่น ทำอย่างนี้ไปเรื่อย ๆ จนได้กราฟลักษณะดังรูป



รูปภาพที่ 2 ภาพแสดงลักษณะกราฟเส้นโค้งรีซีฟเวอร์โอเปอเรติงแคแรกเทอริสติก (Receiver Operating Characteristic Curve: ROC curve)

เมื่อได้กราฟลักษณะดังรูป ต้องนำมาพิจารณาความเหมาะสมในการเลือกค่าขีดแบ่งที่เหมาะสมกับงาน โดยทั่วไปงานปกติ จะเลือกค่าขีดแบ่งที่อยู่ประมาณกึ่งกลาง แต่ในบางกรณีซึ่งมีลักษณะเฉพาะ เช่น การจำแนกผู้ป่วยว่าป่วยเป็นโรคมะเร็งหรือไม่ มักจะเลือกค่าขีดแบ่ง ค่อนไปทางค่าอัตราจริงบวกที่มากขึ้น เนื่องจากต้องการลดการเท็จลบให้น้อยลง กล่าวคือหากผู้ป่วยไม่ได้เป็นโรคมะเร็งแต่ผลการจำแนกบอกว่าเป็น สามารถไปตรวจโรคโดยละเอียดอีกครั้งเพื่อใช้ประกอบการวินิจฉัยโรคเพิ่มเติมได้ แต่ไม่ต้องการให้เกิดกรณีเท็จลบที่ผู้ป่วยป่วยเป็นโรคมะเร็ง แต่ผลการจำแนกบอกว่าไม่ได้เป็น ซึ่งจะส่งผลเสียต่อการรักษา และส่งผลร้ายต่อตัวผู้ป่วยเอง หรือในกรณีการตรวจคุณภาพของข้อมูลผลลัพธ์บนระบบ “ว่าง” ก็ควรเลือกค่าขีดแบ่งค่อนไปทางที่จำแนกว่า ข้อมูลผลลัพธ์มีคุณภาพต่ำ เพราะหากผู้ปฏิบัติงานที่มั่นใจว่าทำถูกและข้อมูลผลลัพธ์มีคุณภาพสูงแต่ถูกจำแนกว่าข้อมูลผลลัพธ์มีคุณภาพต่ำ ก็จะร้องเรียนเข้ามาให้ตรวจสอบได้ แต่หากประเมินผู้ปฏิบัติงานที่ทำข้อมูลผลลัพธ์ออกมาคุณภาพต่ำว่ามีคุณภาพสูง ผู้ปฏิบัติงานคนนั้นก็จะปล่อยผ่านไปเพราะได้

ประโยชน์จากค่าตอบแทน ส่งผลให้ผู้ที่ต้องการข้อมูลได้ข้อมูลผลลัพธ์ที่คุณภาพต่ำกลับไป ซึ่งเป็นสิ่งที่ไม่ต้องการ แต่ก็ไม่ควรเลือกค่าขีดแบ่งที่ค่อนไปทางจำแนกว่าข้อมูลผลลัพธ์มีคุณภาพต่ำมากเกินไป เพราะจะส่งผลให้ผู้ปฏิบัติงานที่ทำข้อมูลคุณภาพสูงออกมารู้สึกไม่พอใจและเลิกทำงานได้ ดังนั้นสุดท้ายก็ควรเลือกค่าขีดแบ่งให้เหมาะสม ไม่มากหรือน้อยเกินไป โดยพิจารณาจากสภาพและลักษณะของงานเป็นหลัก [3]

2.5. การวัดผลแบบจำลองด้วยค่าความแม่น

ค่าความแม่น (Accuracy) คือ อัตราส่วนระหว่างคำตอบที่แบบจำลองตอบถูกกับจำนวนข้อมูลทั้งหมด เป็นการวัดผลค่าแม่นยำแบบทั่วไป ไม่เหมาะแก่การวัดในกรณีที่ข้อมูลกระจายตัวในเหตุการณ์บวกและลบอย่างไม่สมดุลกัน เช่น มีลบมากกว่าบวกเป็นจำนวนมาก เนื่องจากถ้าเป็นแบบนั้น ต่อให้แบบจำลองตอบลบเสมอก็จะส่งผลให้ค่าความแม่นสูงได้ ทั้งที่ความจริงเป็นเพียงการตอบลบเสมอเท่านั้น

ค่าความแม่น เมื่อนำค่าจาก ตารางที่ 1 มาอ้างอิง จะสามารถเขียนสมการได้ดังนี้

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

2.6. การวัดผลแบบจำลองด้วยค่าความเที่ยงและค่าความระลึกได้

2.6.1. ค่าความเที่ยง

ค่าความเที่ยง (Precision) เป็นการคำนึงว่า จากผลลัพธ์ที่แบบจำลองทำนายว่าเป็นบวกทั้งหมด มีกี่อันที่เป็นบวกจริง ๆ

ค่าความเที่ยง เหมาะสำหรับการวัดผลในเหตุการณ์ที่ค่าความเสียหายของเหตุการณ์ เท็จบวก (False Positive) มีค่าสูง เช่น ระบบตรวจจับอีเมลสแปม (Email spam detection) ซึ่งสำหรับระบบนี้ เหตุการณ์เท็จบวก คือ การที่อีเมลฉบับนั้นไม่ใช่สแปม แต่ถูกแบบจำลองจำแนกว่าเป็นสแปม ส่งผลให้อีเมลนั้นหมดความสำคัญและถูกจำแนกไปเป็นสแปม ผู้ใช้งานจะไม่เห็นถึงอีเมลฉบับนั้น หากอีเมลฉบับนั้นสำคัญผู้ใช้งานจะเกิดความเสียหายขึ้น ดังนั้นหากค่าความเที่ยงของระบบตรวจจับอีเมลสแปมมีค่าไม่สูง จะส่งผลให้ผู้ใช้สูญเสียอีเมลสำคัญไปได้

ค่าความเที่ยง เมื่อนำค่าจาก ตารางที่ 1 มาอ้างอิงจะสามารถเขียนสมการได้ดังนี้

$$Precision = \frac{tp}{tp + fp} = \frac{True\ Positive}{Total\ Predicted\ Positive}$$

2.6.2. ค่าความระลึกได้

ค่าความระลึกได้ (Recall) เป็นการคำนึงถึงว่า จากอันที่เป็นบวกจริง ๆ ทั้งหมด แบบจำลองทำนายว่าเป็นบวกกี่อัน

ค่าความระลึกได้ เหมาะสำหรับการวัดผลในเหตุการณ์ที่ค่าความเสียหายของเหตุการณ์ เท็จลบ (False Negative) มีค่าสูง เช่น การตรวจจับกลฉ้อฉล (Fraud detection) ถ้าพฤติกรรมกลฉ้อฉลจริงถูกแบบจำลองจำแนกว่าไม่ใช่กลฉ้อฉล จะส่งผลร้ายต่อธุรกิจของธนาคารได้ หรือการตรวจจับผู้ป่วย (Sick patient detection) หากผู้ป่วยจริงถูกแบบจำลองจำแนกว่าไม่ได้ป่วย ความเสียหายที่เกิดขึ้นอาจรุนแรงได้หากโรคนั้นเป็นโรคติดต่อร้ายแรก

ค่าความระลึกได้ เมื่อนำค่าจาก ตารางที่ 1 มาอ้างอิง จะสามารถเขียนสมการได้ดังนี้

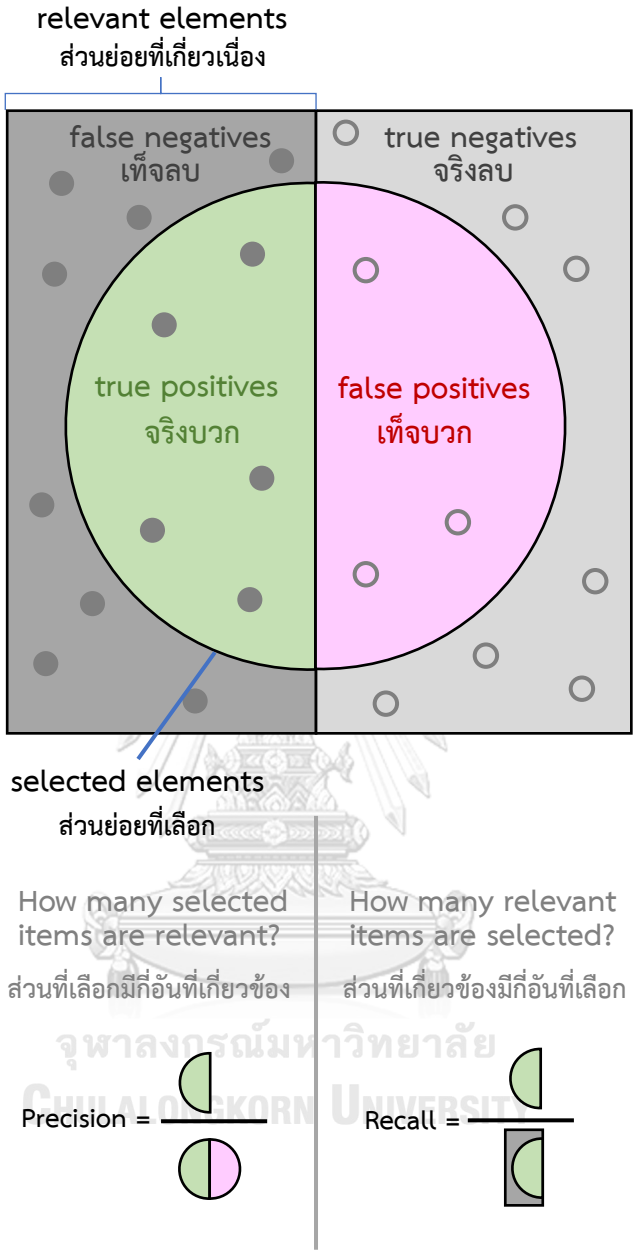
$$Recall = \frac{tp}{tp + fn} = \frac{True\ Positive}{Total\ Actual\ Positive}$$

2.7. การวัดผลแบบจำลองด้วยค่าคะแนนเอฟวัน

คะแนนเอฟวัน (F_1 score) เป็นคะแนนที่สามารถใช้วัดผลได้เมื่อเราต้องการให้ความสำคัญทั้งค่าความเที่ยงและค่าความระลึกได้อย่างสมดุลกัน คะแนนเอฟวันจะวัดผลได้ดีกว่าค่าความแม่น ในกรณีที่เหตุการณ์บวกและลบมีการกระจายตัวอย่างไม่สมดุลกัน เช่น มีเหตุการณ์ลบจริงจำนวนมากกว่าเหตุการณ์บวกจริงเป็นจำนวนมาก เนื่องจากค่าความแม่นจะไม่ได้สนใจว่า เหตุการณ์ฝั่งใดเกิดขึ้นมากกว่า เพียงแค่สามารถทำนายได้ถูกเป็นจำนวนมากก็ถือว่าดีแล้ว แตกต่างกับคะแนนเอฟวันที่มีการคำนึงถึงการทำนายเหตุการณ์ฝั่งบวกและลบด้วย

คะแนนเอฟวัน เป็นการให้คะแนนจากการนำค่าความเที่ยงและค่าความระลึกได้มาคำนวณร่วมกัน สามารถเขียนเป็นสมการได้ดังนี้

$$F_1 = \left(\frac{recall^{-1} + precision^{-1}}{2}\right)^{-1} = \frac{2}{\frac{1}{recall} + \frac{1}{precision}} = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$



รูปภาพที่ 3 ภาพแสดงเหตุการณ์ความสัมพันธ์ระหว่างค่าความเที่ยงและค่าความระลึกได้

บทที่ 3 งานวิจัยที่เกี่ยวข้อง

3.1. คราวด์ซอร์สซิง (Crowdsourcing)

ไม่เพียงแต่จำนวนประชากรผู้ใช้อินเทอร์เน็ตที่เจริญเติบโตขึ้นอย่างรวดเร็ว เนื้อหาและความรู้ต่าง ๆ ที่ประชากรอินเทอร์เน็ตแลกเปลี่ยนความรู้กันก็เพิ่มขึ้นด้วย การที่ผู้ใช้งานอินเทอร์เน็ตมี **ปฏิสัมพันธ์ (interaction)** กันผ่านโลกออนไลน์มากขึ้นเป็นจุดสำคัญที่มีผู้เล็งเห็น และใช้ประโยชน์จากจำนวนคนมหาศาลบนโลกอินเทอร์เน็ต โดยการสร้างสิ่งที่เรียกว่า ระบบคราวด์ซอร์สซิงขึ้นมา เป็นการให้ผู้ใช้งานอินเทอร์เน็ตจำนวนมากเข้ามารวมกันช่วยทำสิ่งต่าง ๆ ตามแต่ละบริการให้สำเร็จผ่านคราวด์ซอร์สซิงแพลตฟอร์ม โดยมีงานและบริการรูปแบบต่าง ๆ มากมายให้คนเข้ามาช่วยกันทำ เช่น การทำงานขนาดเล็ก ซึ่งเป็นงานอย่างง่าย การระดมทุน การสร้างเนื้อหา (content generate) การพัฒนาโปรแกรม (program developing) การออกแบบกราฟิก (graphic design) การแปลภาษา (translation) การทบทวนและทดสอบผลิตภัณฑ์ (product reviews and testing) [1] ซึ่งแต่ละคราวด์ซอร์สซิงแพลตฟอร์มอาจจะเน้นหลักไปที่การให้บริการหนึ่ง ๆ หรือมีหลายบริการรวมอยู่ด้วยกันก็ได้ ปัจจุบันมีคราวด์ซอร์สซิงแพลตฟอร์มเกิดขึ้นหลายแห่งทั่วโลก เช่น microWorkers, clickworker, figure-eight (crowdflower) mturk (Amazon Mechanical Turk) kickstarter Indiegogo โดยแต่ละแพลตฟอร์มมักจะมีลักษณะเด่นที่แตกต่างกันออกไป

คราวด์ซอร์สซิงแพลตฟอร์มที่นักวิจัยและผู้ที่ทำงานด้านการเรียนรู้ของเครื่องใช้กันอย่างแพร่หลายที่สุด คือ mturk ที่ประเทศสหรัฐอเมริกา โดยส่วนมากใช้ในการเก็บรวบรวมข้อมูลเพื่อเป็นข้อมูลตั้งต้นในสำหรับการฝึกฝนเพื่อสร้างแบบจำลองต่าง ๆ ของการเรียนรู้ของเครื่อง การเรียนรู้ของเครื่องจะไม่สามารถทำงานได้หากปราศจากข้อมูลที่มากเพียงพอ ยิ่งข้อมูลมาก แบบจำลองก็จะยิ่งมีความแม่นยำมากขึ้น โดยเฉพาะการเรียนรู้ของเครื่องที่สร้างขึ้นด้วยกรรมวิธี**การเรียนรู้เชิงลึก (Deep Learning)** ดังนั้นข้อมูลจึงเป็นสิ่งที่สำคัญที่สุดในการทำงานด้านการเรียนรู้ของเครื่อง แต่ผู้ปฏิบัติงานบนคราวด์ซอร์สซิงแพลตฟอร์มสามารถเป็นใครก็ได้บนอินเทอร์เน็ต ส่งผลให้เราไม่อาจรับประกันได้ว่าคุณภาพของข้อมูลผลลัพธ์ที่ได้ไปจากคราวด์ซอร์สซิงแพลตฟอร์มจะเป็นไปตามต้องการ ดังนั้นจึงมีความพยายามในการตรวจสอบคุณภาพของข้อมูลผลลัพธ์ที่ได้จากคราวด์ซอร์สซิงแพลตฟอร์ม เพื่อตรวจจับข้อมูลผลลัพธ์ที่มีคุณภาพต่ำกว่าความต้องการและคัดกรองข้อมูลผลลัพธ์นั้นออกไป จะได้ไม่ส่งผลกระทบต่อการสร้างแบบจำลองการเรียนเรียนรู้ของเครื่อง

จากการที่มีความพยายามในการตรวจสอบคุณภาพของข้อมูลผลลัพธ์ซึ่งเป็นคำตอบที่ได้จากคราวด์ซอร์สซิงแพลตฟอร์ม เพื่อตรวจหาข้อมูลผลลัพธ์ที่มีคุณภาพต่ำและคัดกรองออกไป จะทำให้การเรียนรู้ของเครื่องที่นำข้อมูลนี้ไปใช้ทำงานได้ดีขึ้น จึงเป็นที่มาของงานวิจัยฉบับนี้ที่พยายามจะแก้ปัญหาดังกล่าว รวมถึงยังแสดงถึงความสำคัญของปัญหาที่งานวิจัยนี้พยายามจะแก้ไขได้เป็นอย่างดี

3.2. งานวิจัยการตรวจจับคุณภาพข้อมูลผลลัพธ์ด้วยวิธีการทำซ้ำ (Redundant) วิธีนับเสียงข้างมาก (Majority vote) วิธีการสร้างมาตรฐานทองคำ (Gold Standard) และวิธีการใช้ข้อมูลแวดล้อมบางประการช่วยประกอบการตัดสิน

จากความพยายามที่จะตรวจจับข้อมูลผลลัพธ์ที่มีคุณภาพต่ำจึงเริ่มมีผู้วิจัยเกี่ยวกับการตรวจสอบคุณภาพของข้อมูลผลลัพธ์ตั้งแต่ช่วงเริ่มแรกที่คราวด์ซอร์สซิงแพลตฟอร์มได้เกิดขึ้น โดยเบื้องต้นยังใช้ลักษณะวิธีการของการเก็บข้อมูลผลลัพธ์มากกว่าหนึ่งผลในหนึ่งงานหรือที่เรียกว่าวิธีการทำซ้ำ ซึ่งทำให้ต้องเสียงบประมาณในการเก็บข้อมูลมากขึ้น จึงมีงานวิจัย [4] ที่เสนอวิธีการทำซ้ำแบบใหม่ขึ้น โดยได้ทำการเปรียบเทียบการทำซ้ำสองแบบเพื่อพยายามหาวิธีในการลดค่าใช้จ่ายในการเก็บข้อมูล กรรมวิธีที่เสนอคือ 1. ใช้การตัดสินใจจากเสียงส่วนใหญ่ (Majority Decision: MD) 2. การใช้กลุ่มควบคุม (Control Group: CG) ทั้งสองวิธีแตกต่างกันตรงที่ วิธีที่ 1 การตัดสินใจจากเสียงส่วนใหญ่ คือการที่งานหนึ่ง ๆ จะเก็บข้อมูลจากผู้ปฏิบัติงานมากกว่า 1 คน โดยเก็บข้อมูลผลลัพธ์มาจำนวน m ผล เมื่อ m เป็นจำนวนคี่ หลังจากนั้นดูเสี่ยงส่วนใหญ่ว่าในกลุ่ม m ผลนี้ ให้ข้อมูลผลลัพธ์กลับมาว่าเป็นแบบใดมากที่สุด ก็จะสรุปว่าผลนั้นเป็นข้อมูลผลลัพธ์สุดท้าย ส่วนวิธีที่ 2 การใช้กลุ่มควบคุม คือการที่งานหนึ่ง ๆ จะเก็บข้อมูลจากผู้ปฏิบัติงานเพียงคนเดียวเท่านั้น แล้วใช้วิธีการเพิ่มการตรวจสอบผลลัพธ์เข้าไป โดยสร้างงานใหม่ที่เป็นงานตรวจสอบผลลัพธ์ขึ้นมา และให้ผู้ปฏิบัติงานเข้ามาตรวจสอบผลลัพธ์เป็นจำนวน m คน เมื่อ m เป็นจำนวนคี่ งานนี้เป็นเพียงงานที่ให้ผู้ปฏิบัติงานเข้ามาบอกว่า ผลลัพธ์ที่ได้จากผู้ปฏิบัติงานคนแรก เป็นข้อมูลที่เชื่อถือได้หรือไม่ ซึ่งนับเป็นงานอย่างง่าย จึงสามารถจ่ายค่าตอบแทนในจำนวนที่น้อยกว่าค่าตอบแทนที่จ่ายให้ในงานหลักได้ โดยผลการพิสูจน์พบว่า หากแบ่งงานออกเป็น 3 กลุ่ม ได้แก่ 1. งานรูทีน (routine task) หรืองานซึ่งเป็นกิจวัตร 2. งานซับซ้อน (complex task) 3. งานสร้างสรรค์ (creative task) โดยค่าตอบแทนแปรผันตามความยากของงาน กล่าวคือ ค่าตอบแทนน้อยไปมากตามลำดับ วิธีที่ 1 การตัดสินใจจากเสียงส่วนใหญ่จะเหมาะกับงานรูทีนซึ่งเป็นงานที่ง่าย และมีค่าตอบแทนที่ต่ำอยู่แล้ว ในขณะที่วิธีที่ 2 การใช้กลุ่มควบคุม เหมาะสำหรับงานซับซ้อนและงานสร้างสรรค์ ซึ่งมีค่าตอบแทนงานหลักค่อนข้างสูง การสร้างงานย่อยที่ง่ายกว่าเพื่อใช้ตรวจสอบผลจึงเสียค่าตอบแทนที่ต่ำกว่า

นอกจากนี้ ยังมีงานวิจัย [5] ได้ทำการเสนอแนวทางการตรวจสอบคุณภาพของข้อมูลผลลัพธ์ด้วยวิธีการเพิ่มเติม 3 ประการ ได้แก่

1. ใช้ข้อมูลมาตรฐานทองคำ (Gold standard data) เป็นการใช้ข้อมูลที่ทราบอยู่แล้วจากกลุ่มผู้ปฏิบัติงานที่เชื่อถือได้เป็น**พื้นฐานความจริง (ground truth)** แทรกเข้าไปในงานเพื่อเปรียบเทียบกับข้อมูลผลลัพธ์ที่ได้จากผู้ปฏิบัติงานใหม่ หากมีข้อมูลผลลัพธ์ที่ได้มาใหม่ในส่วนที่เรามีข้อมูลผลลัพธ์จากพื้นฐานความจริงอยู่แล้ว มีความแตกต่างไปจากที่เราทราบมากเกินไป ก็มีแนวโน้มว่า ข้อมูลผลลัพธ์อื่นที่ได้จากการปฏิบัติงานในงานเดียวกันนี้จะมีคุณภาพต่ำไปด้วย

2. เมทริกซ์จากภายใน (Intrinsic matric) เป็นการตรวจสอบข้อมูลผลลัพธ์ที่ได้โดยใช้คุณสมบัติการถ่ายทอดบางประการในการทวนสอบข้อมูลผลลัพธ์ด้วยกันเอง เช่น หากผู้ปฏิบัติงานให้ข้อมูลว่า ชอบเสียงที่ได้จากแบบจำลอง A มากกว่าแบบจำลอง B และชอบเสียงที่ได้จากแบบจำลอง B มากกว่าแบบจำลอง C หมายความว่า ในข้อที่ถามผู้ปฏิบัติงานถึงการให้เปรียบเทียบความชอบระหว่างเสียงที่ได้จากแบบจำลอง A กับแบบจำลอง C ผู้ปฏิบัติงานควรจะต้องบอกว่าชอบเสียงที่ได้จากแบบจำลอง A มากกว่าแบบจำลอง C ด้วย ไม่เช่นนั้นอาจแปลได้ว่า ข้อมูลผลลัพธ์ที่ได้มานั้นไม่มีคุณภาพ เพราะมีความขัดแย้งภายในตัวเอง

3. การใช้ข้อมูลเพิ่มเติม (Additional information) เป็นการใช้ข้อมูลอื่นที่ได้จากการปฏิบัติงานในงานนั้นมาประกอบการคัดกรอง เช่น งานที่ให้เปรียบเทียบความชอบระหว่างเสียงที่ได้จากแบบจำลองสองแบบจำลอง ผู้ปฏิบัติงานจะต้องเปิด**สิ่งบันทึกเสียง (sound clip)** หรือคลิปเสียงที่ได้จากแบบจำลองแรกและสิ่งบันทึกเสียงที่ได้จากแบบจำลองที่สองเพื่อฟังและเปรียบเทียบ ดังนั้นหากผู้ปฏิบัติงานให้ข้อมูลผลลัพธ์มาว่าชอบเสียงจากแบบจำลองไหน ด้วยระยะเวลาในการปฏิบัติงานที่น้อยกว่าระยะเวลาของสิ่งบันทึกเสียงทั้งสองรวมกัน อาจหมายความว่าผู้ปฏิบัติงานไม่ได้เปิดสิ่งบันทึกเสียงทั้งสองฟังจริง ๆ ดังนั้นข้อมูลผลลัพธ์ความชอบที่ได้มาจึงไม่มีคุณภาพและควรคัดกรองออกไป

โดยหลังจากที่ได้นำวิธีการทั้งสามไปทดสอบจริง ในการเก็บข้อมูลสำหรับงาน**ระบบข้อความเป็นเสียงพูด (Text to Speech System: TTS system)** กับ 127 ผู้ปฏิบัติงานจริงบน mturk พบว่าวิธีการดังกล่าวสามารถใช้ในการตรวจสอบข้อมูลผลลัพธ์จากผู้ปฏิบัติงานที่มีคุณภาพต่ำออกไปได้จริง แต่ยังมีความเสี่ยงบ้างตรงที่บางครั้งอาจมีการคัดกรองข้อมูลผลลัพธ์จากผู้ปฏิบัติงานที่มีคุณภาพจริงออกไปด้วย

การตรวจสอบคุณภาพด้วยวิธีการข้างต้น มีบางขั้นตอนของการตรวจสอบที่ส่งผลให้ต้องมีการใช้งบประมาณที่มากกว่าปกติในการตรวจสอบ หรือมีการสูญเสียทรัพยากรข้อมูลผลลัพธ์ไปเพื่อใช้ในการตรวจคุณภาพ ซึ่งสำหรับงานการเรียนรู้ของเครื่องที่ต้องเก็บข้อมูลปริมาณมากแล้ว จะยิ่งทำให้ต้องสูญเสียทรัพยากรเงินและข้อมูลไปเป็นจำนวนมาก นับเป็นปัญหาที่เป็นผลพวงมาจากการพยายามตรวจสอบคุณภาพของข้อมูลผลลัพธ์ จึงส่งผลให้งานวิจัยฉบับนี้เล็งเห็นถึงปัญหาดังกล่าวและพยายามจะวิจัยเพื่อเสนอแนวทางอื่น ที่สามารถลดปริมาณการสูญเสียข้อมูลผลลัพธ์ รวมถึงลดปริมาณงบประมาณที่ต้องใช้ในการตรวจคุณภาพของข้อมูลคำตอบลง แต่ยังสามารถตรวจสอบคุณภาพได้อยู่

3.3. งานวิจัยที่ใช้ข้อมูลแวดล้อมจากการปฏิบัติงานในการตรวจสอบคุณภาพเพียงอย่างเดียว

จะเห็นได้ว่า งานวิจัย [5] ได้เริ่มมีการใช้ข้อมูลอื่นนอกเหนือจากการทำซ้ำของข้อมูล การนับเสียงข้างมากและการแทรกมาตรฐานทองคำลงไปในงาน เช่น การใช้เวลาในการทำงาน เข้ามาช่วย

ประกอบการตัดสินใจในการประเมินคุณภาพของข้อมูลผลลัพธ์ ซึ่งเป็นจุดเริ่มต้นของแนวคิดการเปลี่ยนมาใช้ข้อมูลแวดล้อมอื่น ๆ รวมถึงพฤติกรรมการปฏิบัติงานของผู้ปฏิบัติงานเข้ามาช่วยในการตรวจสอบคุณภาพของข้อมูลผลลัพธ์ เพราะจะทำให้ผู้ร้องขอข้อมูลประหยัดงบประมาณลงได้ และได้ใช้ทรัพยากรการทำงานจากผู้ปฏิบัติงานได้เกิดประสิทธิภาพโดยได้ข้อมูลผลลัพธ์มากที่สุด

งานวิจัย [6] เป็นงานที่หันมาใช้ข้อมูลแวดล้อมจากการทำงานเท่านั้นในการประกอบการตัดสินใจ ไม่ได้ใช้การทำซ้ำหรือมาตรฐานทองคำอีก โดยในงานวิจัยนี้ได้ใช้ข้อมูล 3 ประการ ได้แก่ 1. เวลาในการทำงานจนเสร็จ (task completion time) คือ เวลาในการปฏิบัติงานนั้นทั้งหมด ว่าเหมาะสมกับความยาวของงานและสัมพันธ์กับระยะเวลาจริงที่ควรใช้หรือไม่ 2. เวลาในการทำงานแต่ละขั้นที่แตกต่างกัน (difference working phase time) คือ เวลาในการปฏิบัติงานแต่ละส่วน เช่น เวลาในการอ่านคำสั่ง เวลาในการปฏิบัติงานส่วนย่อย ๆ ควรจะมีระยะเวลาที่สมเหตุสมผล และ 3. เวลาในการพิจาณา (consideration time) นับตั้งแต่เวลาที่ผู้ปฏิบัติงานได้เห็นคำถามหรือคำสั่งจนกระทั่งผู้ปฏิบัติงานตัดสินใจส่งคำตอบข้อมูลผลลัพธ์สุดท้ายเสร็จสิ้น เมื่อเก็บข้อมูลทั้ง 3 อย่างนี้ในแต่ละการปฏิบัติงานได้แล้ว ก็นำมาสร้างแบบจำลอง**ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM)** เพื่อใช้ในการจำแนกข้อมูลผลลัพธ์การปฏิบัติงานว่ามีคุณภาพหรือไม่ โดยจากการทดสอบบนงานทดสอบภาษาอังกฤษ 25 คำถาม กับผู้ปฏิบัติงานรวม 215 คน พบว่า แบบจำลองที่สร้างได้มีความแม่นยำในการจำแนกถึง 88.67%

งานวิจัยในส่วนนี้ ได้จุดประกายแนวคิดและได้เสนอให้เห็นถึงแนวทางอื่นที่สามารถตรวจสอบคุณภาพของข้อมูลผลลัพธ์ได้โดยไม่ต้องสูญเสียทรัพยากรอันมีค่าไปแบบงานวิจัยก่อน ๆ รวมถึงยังได้แสดงให้เห็นถึงการใช้ลักษณะพฤติกรรมในการทำงานบางประการในการตรวจสอบคุณภาพของข้อมูลได้อย่างเป็นเหตุเป็นผล จึงเป็นจุดเริ่มต้นให้งานวิจัยฉบับนี้ ได้แนวคิดในการใช้ข้อมูลแวดล้อมอย่างพฤติกรรมในการทำงานของผู้ปฏิบัติงานเพื่อตรวจสอบคุณภาพของข้อมูล ซึ่งสามารถช่วยลดปัญหาการสูญเสียทรัพยากรเงินและข้อมูลลงได้

3.4. งานวิจัยที่สะท้อนว่าการขยับเมาส์ของผู้ใช้งานบนแพลตฟอร์มสามารถบ่งบอกถึงพฤติกรรมอื่นของผู้ใช้งานได้

เมื่อเริ่มมีการนำข้อมูลแวดล้อม เช่น เวลาจากการปฏิบัติงาน มาใช้ในการตัดสินใจ จึงเริ่มมีผู้ที่เห็นว่า พฤติกรรมอื่นของผู้ปฏิบัติงานจากกระบวนการทำงาน น่าจะสามารถบ่งบอกถึงลักษณะในการทำงานของผู้ปฏิบัติงาน ซึ่งสะท้อนถึงคุณภาพของข้อมูลผลลัพธ์ที่ได้จากการปฏิบัติงานเช่นกัน โดยจากงานวิจัย [7] สามารถสะท้อนให้เห็นได้ว่า **การเคลื่อนที่ของเมาส์ (mouse movement)** มี **ความสัมพันธ์ (correlation)** กับ**การเคลื่อนที่ของสายตา (eye movement)** อย่างมีนัยสำคัญ แสดงให้เห็นว่า หากเราติดตามการเคลื่อนที่ของเมาส์ จะสามารถวิเคราะห์ออกมาได้ว่า ณ ขณะหนึ่ง

ๆ ผู้ปฏิบัติงานเพ่งความสนใจ หรือใช้สายตามองอยู่ ณ ตำแหน่งใด ซึ่งสามารถนำสิ่งนี้ไปวิเคราะห์ลักษณะการทำงานได้เพิ่มมากขึ้น

งานวิจัยข้างต้นสามารถยืนยันแนวคิดของงานวิจัยฉบับนี้ได้ว่า ข้อมูลที่เราสามารถเก็บได้จากแพลตฟอร์ม สามารถสะท้อนถึงพฤติกรรมในการทำงานของผู้ปฏิบัติงานได้จริง

3.5. งานวิจัยที่ใช้การตรวจจับพฤติกรรมในการทำงาน

งานวิจัย [3], [8], [9] เห็นว่าการที่หันมาใช้ข้อมูลแวดล้อมอื่นในการตรวจสอบคุณภาพของข้อมูลจะทำให้สามารถประหยัดค่าใช้จ่ายลงได้ รวมถึงเห็นว่า ข้อมูลอื่นที่สามารถเก็บมาได้เวลาที่ผู้ปฏิบัติงานทำงาน เช่น การเคลื่อนที่ของเมาส์ สามารถสะท้อนพฤติกรรมการทำงานของผู้ปฏิบัติงานได้ด้วย จึงได้ทำการเก็บข้อมูลพฤติกรรมการทำงาน 10 อย่าง ได้แก่

1.เวลาก่อนเริ่มทำงาน (startup time) คือ ช่วงเวลาที่ผู้ปฏิบัติงานเปิดงานขึ้นมาแล้วแต่ยังไม่ขยับเมาส์ เพียงแค่กวาดสายตาผ่านเนื้อหา

2. จำนวนการเคลื่อนที่ย่อย (sub movement count) คือ จำนวนการเคลื่อนที่ย่อยของเมาส์ในแต่ละคำถาม

3. จำนวนการเคลื่อนที่ย่อยทั้งหมด (overall sub movement count) คือ จำนวนการเคลื่อนที่ย่อยของเมาส์ในทุกคำถามทั้งหมดในงานรวมกัน

4. จำนวนการหยุดทั้งหมด (overall number of pause) คือ จำนวนมากหยุดเมาส์ทั้งหมด

5. ค่ามัธยฐานเวลาหยุดทั้งหมด (overall median pause duration) คือ ค่ามัธยฐานของเวลาในการหยุดเมาส์ทั้งหมด

6. จำนวนการคลิกพิเศษ (number of extra click) คือ จำนวนการคลิกเมาส์ที่พิเศษขึ้นมาโดยไม่ใช้เป็นการคลิกเพื่อตอบคำถาม

7. ค่ามัธยฐานเวลาระหว่างคำถาม (median inter-question time) คือ ค่ามัธยฐานของเวลาระหว่างแต่ละคำถาม

8. ค่ามัธยฐานเวลาระหว่างการเคลื่อนที่ย่อย (median inter-question sub movement) คือ ค่ามัธยฐานของเวลาระหว่างการเคลื่อนที่ย่อยของเมาส์แต่ละ

9. ค่ามัธยฐานความเร็วตัวชี้ตำแหน่ง (median cursor speed) คือ ค่ามัธยฐานของความเร็วของการเคลื่อนที่ของเมาส์หรือ**ตัวชี้ตำแหน่ง (cursor)**

10. ค่ามัธยฐานความเร่งตัวชี้ตำแหน่ง (median cursor acceleration) คือ ค่ามัธยฐานของความเร่งของการเคลื่อนที่ของเมาส์หรือ**ตัวชี้ตำแหน่ง (cursor)**

งานวิจัยนี้ได้เก็บข้อมูลทั้ง 10 ค่าข้างต้น จากการให้ผู้ปฏิบัติงานทั้งหมด 172 คน ทำงานประเมินคุณภาพวิดีทัศน์แบบปรับได้ (Adaptive video quality assessment) เป็นการให้

ผู้ปฏิบัติงานดูวิดีทัศน์ 4 อัน อันละ 60 วินาที ที่แต่ละวิดีทัศน์มี**อัตราบิต (bit rate)** ในการแสดงผลที่แตกต่างกัน หลังจากนั้นให้ผู้ปฏิบัติงานตอบคำถาม 16 ข้อ ซึ่งเป็นการประเมินวิดีทัศน์ในแต่ละแง่มุม โดยมีวิธีการให้ผู้ปฏิบัติงานให้คะแนนวิดีทัศน์ด้วยการให้คะแนนแบบ **มาตราส่วนไลเคิร์ต 5 คะแนน (5 points likert scale)** ที่แตกต่างกันในแต่ละคนโดยสุ่ม มีทั้งหมด 4 รูปแบบ ได้แก่ 1. ปุ่มวิทยุ (radio button) 2. แถบเลื่อน (slide bar) 3. ฟิลด์ใส่หมายเลข (number field) 4. ดาว (star) หลังจากนั้นนำข้อมูลที่ได้ไปสร้างแบบจำลองที่ใช้ในการตรวจจับข้อมูลผลลัพธ์ที่มีคุณภาพต่ำด้วย **แบบจำลองนาอีฟเบย์หลายกลุ่ม (Multiclass Naïve Bayes Model)** โดยสุ่มข้อมูลบางส่วนมากสร้างแบบจำลองที่แตกต่างกันทั้งหมด 5 แบบ เพื่อสำรวจว่าลักษณะข้อมูลใดที่มีผลต่อคุณภาพของข้อมูลมากที่สุด ผลการทดลองปรากฏว่า ลักษณะที่ส่งผลกระทบต่อการทำนายมากที่สุด ได้แก่ 1. จำนวนการเคลื่อนที่ย่อย (sub movement count) 2. จำนวนการคลิกพิเศษ (number of extra click) 3. ค่ามัธยฐานเวลาระหว่างคำถาม (median inter-question time) 4. ค่ามัธยฐานความเร็วตัวชี้ตำแหน่ง (median cursor speed) และรูปแบบการให้คะแนนที่แบบจำลองสามารถทำนายได้ผลแม่นยำมากที่สุดคือ 1. การให้คะแนนแบบปุ่มวิทยุ และ 2. การให้คะแนนแบบดาว ซึ่งการทำนายของแบบจำลองยังให้ผลไม่เป็นที่น่าพอใจ เพราะเกิดการ**ประเมินคุณภาพสูงไป (overestimate)** กล่าวคือ แม้แต่งานที่มีคุณภาพต่ำบางงานก็ประเมินว่ามีคุณภาพสูง ซึ่งเป็นสิ่งที่ไม่ต้องการ ท้ายที่สุด งานวิจัยนี้จึงได้รวมการประเมินทุกแบบจำลองแล้วเลือกผลสุดท้ายที่ประเมินต่ำที่สุด จึงได้ผลความแม่นยำอยู่ที่ 82.6% จะเห็นได้ว่าแบบจำลองที่สร้างขึ้น ยังไม่สามารถใช้ได้ดีในทุกรูปแบบคำถาม แม้แต่งานเดียวกันแค่เปลี่ยนลักษณะการตอบคำถามก็ส่งผลต่อความแม่นยำของแบบจำลองได้

นอกจากนี้ยังมีงานวิจัยอื่นที่หันมาใช้พฤติกรรมในการปฏิบัติงานเพื่อสร้างแบบจำลองตรวจจับคุณภาพของข้อมูลผลลัพธ์อื่นอีก อย่างงานวิจัยของ [10] ใช้การเก็บข้อมูลจาก**ชั้นแอปพลิเคชัน (application layer)** บน**เว็บเบราว์เซอร์ (web browser)** เช่น **การเคลื่อนที่ของเมาส์ (mouse movement) การคลิก (click) การเลื่อน (scrolling) การเปลี่ยนขนาดหน้าต่าง (window resizing) เหตุการณ์คัดลอก-วาง (copy-paste events) การได้โฟกัสและสูญเสียโฟกัสจากหน้าต่าง (window lost/got focus)** โดยจากข้อมูลเหล่านี้ งานวิจัยนี้ได้สกัดข้อมูลออกมาทั้งหมด 28 ลักษณะ โดยมี 22 ลักษณะที่ใช้ประกอบกับค่าทางสถิติ 7 อย่าง ได้แก่ 1. ยอดรวม (Total) 2. ค่าเบี่ยงเบนมาตรฐาน (Standard Deviation: SD) 3. ค่าเอนโทรปีของแชนนอน (Shannon's Entropy: SE) 4. ค่าต่ำสุด (Minimum: Min) 5. ค่าสูงสุด (Maximum: Max) 6. ค่าเฉลี่ย (Average: Avg) 7. ค่ามัธยฐาน (Median: Med) รวมได้ลักษณะทั้งหมด (22 * 7) + 6 = 160 ลักษณะ ดังตารางที่ 2 โดยนำข้อมูลจริงที่เก็บได้จากผู้ปฏิบัติงานจริงบนคราวด์ซอร์สซิงแพลตฟอร์มทั้งหมดมาเป็นข้อมูลฝึกฝนเพื่อใช้ในสร้างแบบจำลองตรวจสอบคุณภาพของข้อมูลผลลัพธ์

ด้วยกรรมวิธีแบบ **เกรเดียนต์บูสต์ดีซิชันทรี (Gradient Boosted Decision Tree: GBDT)** ซึ่งแตกต่างกับงานอื่นที่ใช้ข้อมูลฝึกฝนจากกลุ่มควบคุมตัวอย่างซึ่งเป็นมาตรฐานทองคำมาใช้ในการฝึกฝนเท่านั้น และได้ผลสรุปว่าแบบจำลองสามารถคัดแยกคุณภาพได้ดีกว่าวิธีการใช้มาตรฐานทองคำปกติและดีกว่าแบบจำลองที่ใช้ข้อมูลฝึกฝนจากกลุ่มตัวอย่างเพียงอย่างเดียว

Table 2: Behavioral features used to characterize judges, generated from the logs for the Single (S), Pairwise (P) and List (L) judging tasks. The * means a set of features: Total (Tot), standard deviation (SD), Shannon's entropy (SE), minimum (Min), maximum (Max), average (Avg) and Median (Med). Time is in seconds.

ID	Description
HITs	Total number of HITs completed by judge
Days	Total number of days when judge worked
AvgRPH	Average rate of judgment per hour
JudgmentsMade*	Number of relevance or preference judgments per HIT (e.g., changed their minds)
OptionalRatingsRatio*	Ratio of optional judgments completed per HIT
CmtLen*	Number of characters in comment field per HIT
HITsWithNoCmtRatio	Proportion of judge's HITs with no comment
HitsWithoutURLViewClickRatio	Proportion of judge's HITs when no URLs were clicked
HITsWithOptionalRatingsRatio	Proportion of judge's HITs where judge provided additional optional judgments
TimeSpent*	Number of seconds spent per HIT
SpamTime*	Time spent outside the HIT (excludes time spent on viewing clicked URLs) per HIT
ActiveRatio*	(TimeSpent - SpamTime) / TimeSpent
MouseActivity*	Number of logged mouse move events per HIT
MouseNormActivity*	Number of logged mouse move events per second and per HIT
MouseDwellTime*	Dwell times between mouse move events per HIT
MouseNormDwellTime*	Dwell times between mouse move events per second and per HIT
MicroMouseDwellTime*	Mouse move dwell times over all mouse move events in a judge's HITs (not per HIT)
ClickActivity*	Number of logged mouse clicks per HIT
ClickNormActivity*	Number of logged mouse clicks per second and per HIT
ClickDwellTime*	Dwell times between clicks per HIT
ClickNormDwellTime*	Dwell times between clicks per second and per HIT
MicroClickDwellTime*	Click dwell times over full set of clicks in judge's HITs (not per HIT)
OrientationTime*	Time elapsed before first click per HIT
URLViewClicks*	Number of URLs clicked per HIT
URLViewClickDwellTime*	Time spent on viewing a web page after clicking on its URL per HIT
CopyEvents*	Number of copy-paste events per HIT, e.g., into comment field
ScrollEvents*	Number of scrolling events per HIT
ResizeEvents*	Number of window resize events per HIT

ตารางที่ 2 ตารางแสดงข้อมูลลักษณะพฤติกรรมที่ใช้ตัดสินคุณภาพ ซึ่งเสนอจากงานวิจัยที่เกี่ยวข้อง

ยังมีงานวิจัยอื่นที่สามารถใช้ยืนยันได้ว่าข้อมูลพฤติกรรมที่เก็บได้ระหว่างที่ผู้ปฏิบัติงานทำงานมีผลต่อแบบจำลองคัดกรองคุณภาพจริง โดยงานวิจัย [11] ได้เสนอว่า ข้อมูลพฤติกรรมที่เก็บได้จากการปฏิบัติงานใน**งานอย่างง่าย (simple task)** มีน้อย ส่งผลให้การนำข้อมูลเหล่านั้นมาสร้างแบบจำลองประเมินคุณภาพของข้อมูลผลลัพธ์เป็นไปได้ยาก จึงเสนอให้มีการ**ถ่วงดุล (trade-off)** ระหว่างความแม่นยำในการตรวจสอบคุณภาพของข้อมูลผลลัพธ์กับความซับซ้อนของงาน โดยงานวิจัยนี้ได้เสนอให้เพิ่มความซับซ้อนบางประการลงไปในงานอย่างง่าย เพื่อให้สามารถเก็บข้อมูลการปฏิบัติงานได้เพิ่มมากขึ้น ซึ่งในงานวิจัยนี้ได้ทดสอบในงานของการดูข้อความจาก**ทวีตเตอร์ (Tweeter)** แล้วตอบคำถาม 2 ข้อ ซึ่งในตอนแรก ลักษณะงานจะเป็น มีข้อความทวีตเตอร์ปรากฏขึ้นมาพร้อมคำถาม 2 ข้อที่ต้องการถาม ซึ่งเป็นงานที่ง่าย ผู้ปฏิบัติงานสามารถเสร็จงานได้ในเวลาเฉลี่ยเพียงงานละ 6 วินาทีเท่านั้น งานวิจัยนี้ได้ทำการทดสอบโดยเพิ่มความซับซ้อนของงานเข้าไป

โดยในตอนแรกจะไม่ขึ้นข้อความจากทวีตเตอร์ ขึ้นเพียงคำถาม 2 ข้อ ผู้ปฏิบัติงานจะต้องกดปุ่ม 1 ปุ่มค้างไว้ เพื่อให้ข้อความจากทวีตเตอร์ปรากฏขึ้นมาให้อ่าน ซึ่งทำให้มีขั้นตอนการทำงานที่ซับซ้อนมากขึ้นเล็กน้อย ส่งผลให้ผู้ปฏิบัติงานใช้เวลาเฉลี่ยในการทำงานเพิ่มขึ้นเป็น 10 วินาที จากการเก็บข้อมูลพฤติกรรมการทำงานจากงานทั้งสองแบบมาสร้างแบบจำลองประเมินคุณภาพของข้อมูลผลลัพธ์ด้วย **ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM)** พบว่าจากตอนงานอย่างง่ายแบบจำลองมีความแม่นยำเพียง 47% แต่เมื่อเพิ่มความซับซ้อนของงานเข้าไปแล้วทำให้เก็บข้อมูลมาใช้ในการสร้างแบบจำลองได้มากขึ้น ส่งผลให้ความแม่นยำเพิ่มขึ้นไปถึง 87% จึงเป็นผลที่แสดงให้เห็นอย่างชัดเจนว่าข้อมูลพฤติกรรมที่เก็บได้ระหว่างการทำงานของผู้ปฏิบัติงานส่งผลต่อการสร้างแบบจำลองตรวจสอบคุณภาพของข้อมูลผลลัพธ์เป็นอย่างมาก

จากงานวิจัยในส่วนนี้ แสดงให้เห็นอย่างชัดเจนว่า ข้อมูลพฤติกรรมการทำงานของผู้ปฏิบัติงานสามารถนำมาใช้ในการตรวจจับข้อมูลผลลัพธ์คุณภาพต่ำได้จริง ทั้งข้อมูลการขยับเมาส์ เวลาในการทำงาน รวมถึงลักษณะการทำงานอื่น ๆ โดยยิ่งมีข้อมูลลักษณะการปฏิบัติงานมาก ยิ่งมีแนวโน้มที่จะสร้างแบบจำลองในการตรวจจับข้อมูลผลลัพธ์ที่มีคุณภาพต่ำได้ดียิ่งขึ้น งานวิจัยฉบับนี้จึงจะยึดแนวคิดในการใช้พฤติกรรมการทำงานเพื่อสร้างแบบจำลองตรวจจับคุณภาพ และนำแนวทางนี้มาพัฒนาต่อยอดต่อไป

3.6. งานวิจัยที่จำแนกประเภทลักษณะพฤติกรรมของผู้ปฏิบัติงานซึ่งอธิบายถึงพฤติกรรมที่ส่งผลให้ข้อมูลผลลัพธ์มีคุณภาพต่ำ

งานวิจัย [12] ได้มีการแบ่งผู้ปฏิบัติงานออกเป็น 4 กลุ่ม ได้แก่

1. สแปมเมอร์แบบสุ่ม (Random Spammer) คือ กลุ่มผู้ปฏิบัติงานที่ให้ข้อมูลแบบสุ่มในทุกคำตอบ ตอบคำถามเดาสุ่มอย่างไม่มีแบบแผน

2. สแปมเมอร์แบบเอกรูป (Uniform Spammer) คือ กลุ่มผู้ปฏิบัติงานที่ให้ข้อมูลแบบมีแบบรูปที่ชัดเจนในการตอบ เช่น ตอบตัวเลือก 2 ในทุกคำถามของงาน

3. ผู้ปฏิบัติงานเลินเล่อ (Sloppy Worker) คือ กลุ่มผู้ปฏิบัติงานที่มีความแม่นยำในการให้ข้อมูลต่ำ อาจเป็น**ผู้ประสงค์ร้าย (malicious)** หรือผู้ที่ต้องการ**โกง (cheat)** ซึ่งให้ข้อมูลแบบผิด ๆ พยายามหลบรูปการตตรวจจับคุณภาพ หรืออาจจะเป็นผู้ขาดความรู้ความสามารถ และทักษะในการทำงานนั้นให้สำเร็จก็ได้

4. ผู้ปฏิบัติงานที่เหมาะสม (Proper Worker) คือ กลุ่มผู้ปฏิบัติงานที่สามารถให้ข้อมูลผลลัพธ์ที่มีความแม่นยำได้ตามที่ผู้ที่ต้องการข้อมูลต้องการ

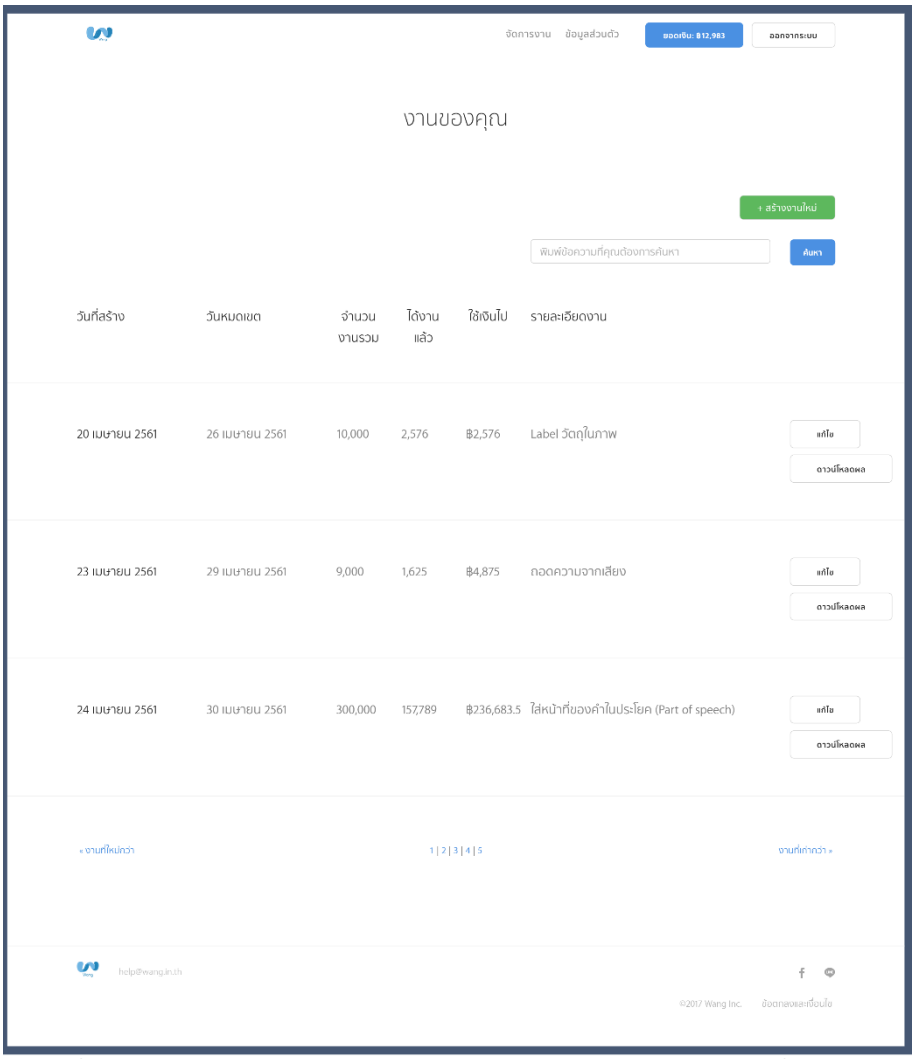
จากข้อมูลในส่วนนี้แสดงให้เห็นว่า เราสามารถแบ่งกลุ่มผู้ปฏิบัติงานได้จากพฤติกรรมการทำงาน เนื่องจากพฤติกรรมการทำงานสามารถถูกจำแนกออกเป็นกลุ่ม ๆ ได้ด้วย**แบบรูป (pattern)**

บางประการ ดังนั้นจึงมีความเป็นไปได้ที่งานวิจัยฉบับนี้จะนำพฤติกรรมการทำงานมาแบ่งเป็นกลุ่ม ประกอบกับความรู้จากงานวิจัยฉบับอื่นที่แสดงให้เห็นว่า พฤติกรรมการทำงานสามารถสะท้อนถึงคุณภาพของข้อมูลผลลัพธ์ได้ ดังนั้นแนวคิดที่จะจำแนกกลุ่มพฤติกรรมการทำงาน ซึ่งสะท้อนถึงการแบ่งกลุ่มข้อมูลผลลัพธ์จากคุณภาพนั้นสามารถทำได้ ส่งผลให้สามารถตรวจหาข้อมูลผลลัพธ์กลุ่มที่มีคุณภาพต่ำได้เช่นกัน

บทที่ 4 แนวคิดและวิธีการดำเนินงานวิจัย

4.1. การเก็บข้อมูลด้วยคราวด์ซอร์สซิงแพลตฟอร์ม “ว่าง”

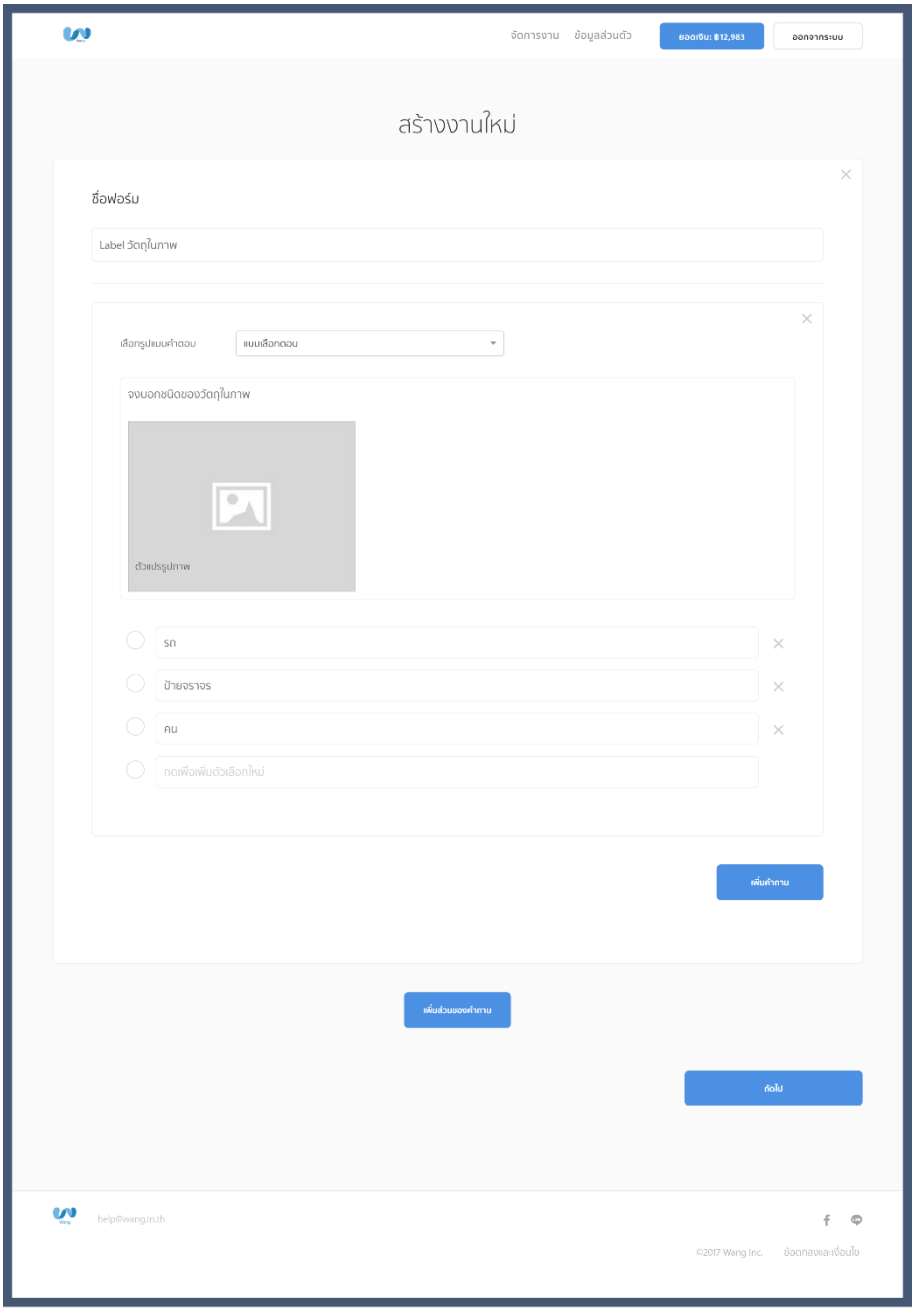
การเก็บข้อมูลด้วยคราวด์ซอร์สซิงแพลตฟอร์ม “ว่าง” เมื่อผู้ที่ต้องการข้อมูลเข้ามาในระบบจะเข้าสู่หน้ารายการงานดัง รูปภาพที่ 4 ซึ่งสามารถบริหารจัดการงานและรายละเอียดของงานได้ตั้งแต่ดูรายการงานของเรา สร้างงานใหม่ แก้ไขงาน ลบงาน และดูข้อมูลผลลัพธ์การตอบของงาน



รูปภาพที่ 4 ภาพแพลตฟอร์ม “ว่าง” ตัวอย่างหน้ารายการงาน สำหรับผู้ที่ต้องการข้อมูล

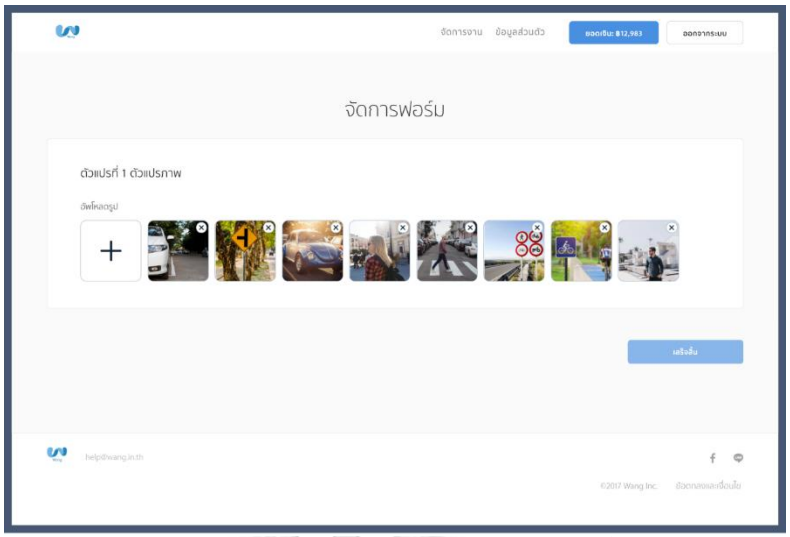
เมื่อผู้ต้องการข้อมูลจะร้องขอข้อมูล ให้กดสร้างงานใหม่เพื่อเข้าสู่หน้าสร้างงาน ซึ่งมีลักษณะเป็นรูปแบบ (form) ให้สร้างงาน กรอกรายละเอียดของงาน ตั้งชื่องาน เพิ่มคำถาม เลือกรูปแบบการตอบคำถาม ตั้งคำถามและอธิบายวิธีการตอบ ซึ่งมีลักษณะการแสดงผลคล้ายหน้าที่ผู้ปฏิบัติงานเห็น

ตอนทำงานจริง นอกจากนี้ ยังสามารถสร้างตัวแปรสำหรับนำข้อมูลบรรจุเข้าไปในตัวแปร ไม่ว่าจะเป็นตัวแปรข้อความ ตัวแปรเสียง ตัวแปรรูปภาพ หรือตัวแปรวิดีโอ เพื่อให้ระบบเปลี่ยนการแสดงผลข้อมูลที่แตกต่างกันในแต่ละรอบที่ทำงานได้โดยอัตโนมัติ



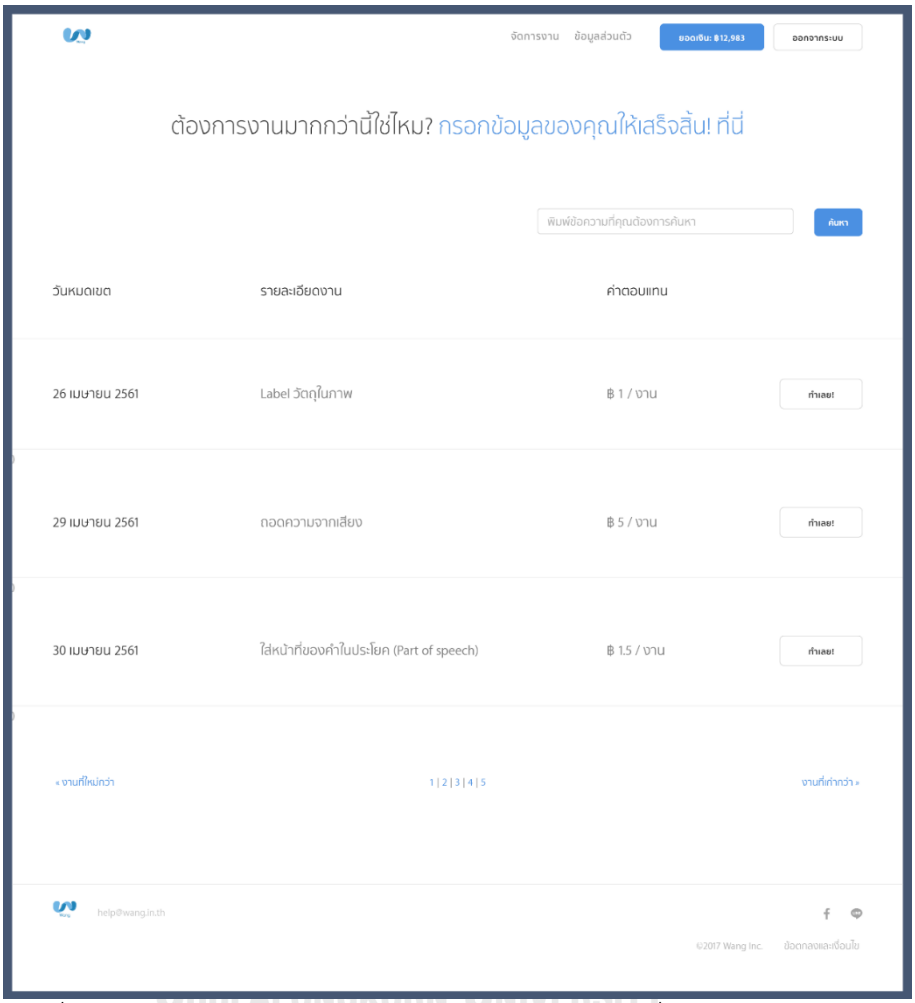
รูปภาพที่ 5 ภาพแพลตฟอร์ม “ว่าง” ตัวอย่างหน้าสร้างงานใหม่

ต่อมาผู้ที่ต้องการข้อมูลจะต้องบรรจุข้อมูลเข้าไปในตัวแปรที่สร้างขึ้น เพื่อให้ระบบสามารถนำข้อมูลเหล่านี้มาแสดงผลในส่วนของตัวแปร เพื่อให้ผู้ปฏิบัติงานทำงานตอบข้อมูลผลลัพธ์ให้กับข้อมูลที่แตกต่างกันไปในแต่ละรอบที่เข้ามาทำงาน



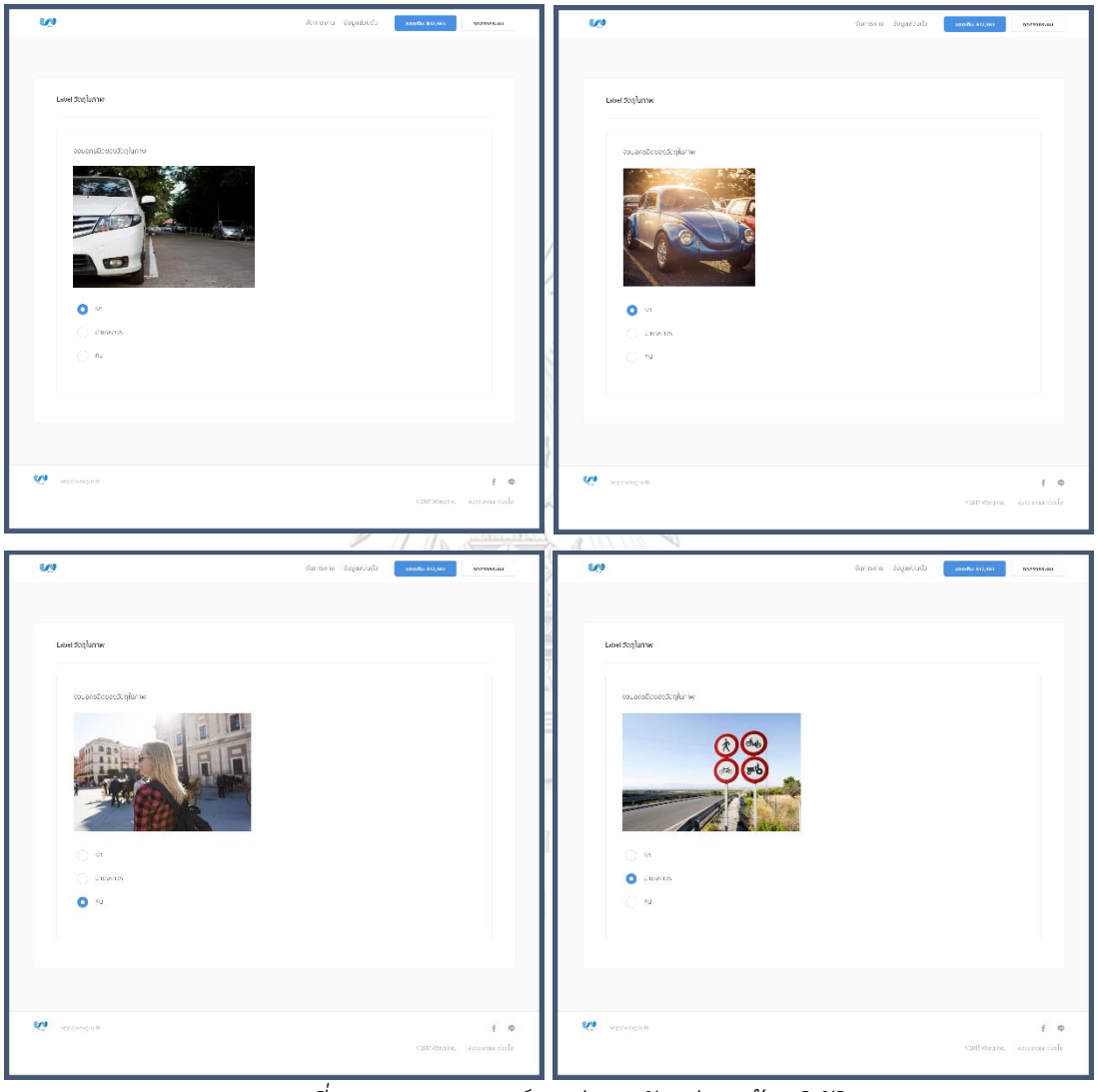
รูปภาพที่ 6 ภาพแพลตฟอร์ม “ว่าง” ตัวอย่างหน้านำภาพบรรจุขึ้น (upload) ตัวแปรภาพ

ส่วนผู้ปฏิบัติงาน เมื่อเข้าสู่ระบบ จะพบกับรายการงานที่สามารถทำได้ พร้อมรายละเอียดค่าตอบแทนของงานให้เลือกกดเข้าไปทำได้ตามต้องการ



รูปภาพที่ 7 ภาพแพลตฟอร์ม “ว่าง” ตัวอย่างหน้ารายการงานที่มีให้ทำ สำหรับผู้ปฏิบัติงาน

เมื่อกดเข้าไปปฏิบัติงาน จะพบกับข้อแนะนำ แนวทางการปฏิบัติงานและคำถามของงานนั้นที่เจ้าของงานสร้างไว้ให้ ก็สามารถปฏิบัติตามข้อแนะนำเพื่อทำงานให้สำเร็จได้ เมื่อทำงานสำเร็จ ก็จะได้รับค่าตอบแทนกลับไปตามที่ถูกกำหนดไว้



รูปภาพที่ 8 ภาพแพลตฟอร์ม “ว่าง” ตัวอย่างหน้าปฏิบัติงาน

4.2. การตรวจสอบคุณภาพของข้อมูลบนคราวด์ซอร์สซิงด้วยการเพิ่มกระบวนการในการเก็บข้อมูล

การเก็บข้อมูลหรือการสร้างฐานข้อมูลโดยทั่วไป ไม่ว่าจะเป็นการลงพื้นที่เก็บข้อมูลด้วยตัวเอง จ้างทีมงานมาช่วยเก็บ หรือการแจกแบบสำรวจ ล้วนเก็บข้อมูลได้ช้า ไม่เพียงพอต่อความต้องการ รวมถึงยังได้ข้อมูลจากกลุ่มประชากรที่ไม่หลากหลาย ซึ่งอาจจะไม่ตอบโจทย์ความต้องการของผู้หาข้อมูล ทั้งในแง่ของการเก็บข้อมูลแบบสอบถามทางสถิติเพื่อทำงานวิจัยต่าง ๆ หรือการสร้างชุดข้อมูลสำหรับใช้ในการสอนแบบจำลองการเรียนรู้ของเครื่อง จึงเป็นที่มาให้คนหันมาเก็บข้อมูลด้วยคราวด์ซอร์สซิงแพลตฟอร์มเพื่อแก้ไขปัญหาดังกล่าว

คราวด์ซอร์สซิงแพลตฟอร์ม สามารถแบ่งออกได้หลากหลายประเภท ในงานนี้จะพูดถึงคราวด์ซอร์สซิงแพลตฟอร์มที่นำมาใช้สำหรับงานเก็บข้อมูลต่าง ๆ คล้าย Amazon Mechanical Turk การเก็บข้อมูลด้วยระบบคราวด์ซอร์สซิง คือ การที่เรานำงานที่ต้องการเก็บข้อมูล มาแบ่งออกเป็นงานขนาดเล็กที่เรียกว่า microtask และกระจายงานเหล่านั้นขึ้นไปบนระบบออนไลน์ เพื่อให้คนจำนวนมากสามารถเข้ามาช่วยกันทำได้โดยง่าย ซึ่งคนที่เข้ามาทำงานเหล่านั้นสามารถเป็นใครก็ได้ ทำงานที่ไหนในโลกก็ได้ โดยจะได้รับค่าตอบแทนกลับไปหลังจากทำงานเสร็จ ด้วยเหตุนี้ จึงทำให้การเก็บข้อมูลปริมาณมากและหลากหลาย เป็นไปได้โดยสะดวกรวดเร็วมากยิ่งขึ้น แต่ในขณะเดียวกันก็ส่งผลให้มีปัญหาในเรื่องคุณภาพของข้อมูลที่ได้รับมา เนื่องจากคนที่เข้ามาช่วยกันทำงานสามารถเป็นใครก็ได้ ซึ่งอาจจะทำไม่ได้ทำตามข้อกำหนดของงาน ตั้งใจโกง หรือมั่วคำตอบ เพื่อให้ได้ค่าตอบแทนมากที่สุดโดยเร็ว อย่างไรก็ตามคุณภาพของข้อมูลนับเป็นสิ่งสำคัญอย่างมากในทุก ๆ งานที่นำข้อมูลไปใช้ต่อ ดังนั้นการตรวจสอบคุณภาพของข้อมูลเพื่อลดปัญหานี้ลงจึงเป็นสิ่งสำคัญอย่างยิ่ง

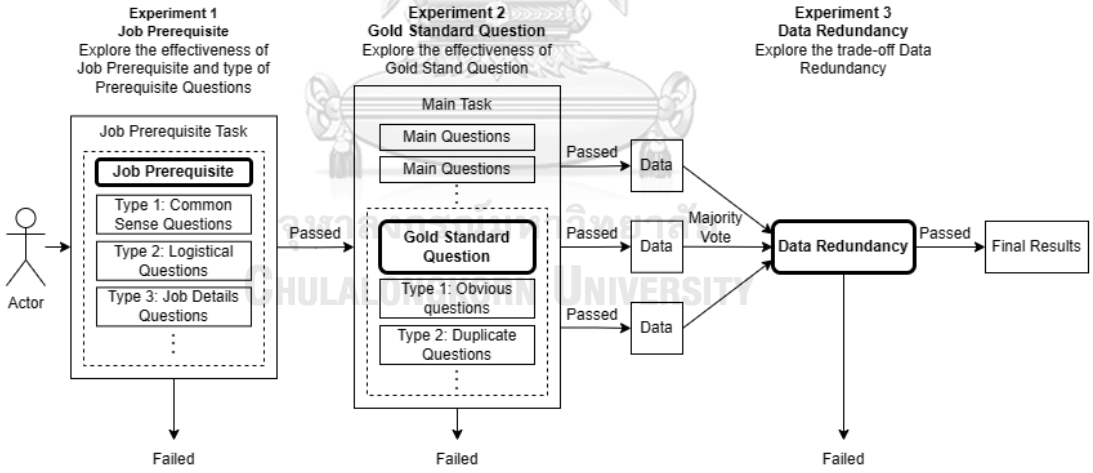
เมื่อผู้ร้องขอข้อมูล (Data Requester) ต้องการจะสร้างงานของตนเองขึ้นบนระบบคราวด์ซอร์สซิงเพื่อเก็บข้อมูล จะประสบกับปัญหาเกี่ยวกับการออกแบบชุดคำถามหรือการสร้างงานเก็บข้อมูลของตนเอง เช่น ควรจะมีแบบทดสอบ (Job Prerequisites) กี่ข้อ คำถามที่ตั้งควรจะเป็นแบบใด จะแทรกคำถามมาตรฐานทองคำ Gold Standard Question ลงไปอย่างไร หรือ ควรจะเก็บการทำซ้ำของข้อมูล (Data Redundancy) จำนวนกี่ครั้ง จากงานวิจัยที่เกี่ยวข้องพบว่า การออกแบบงานและกระบวนการเก็บข้อมูล หากเพิ่มคำถามหรือบททดสอบต่าง ๆ เข้าไปยิ่งมาก ก็จะสามารถคัดกรองคนทำงาน และข้อมูลที่มีคุณภาพต่ำออกได้มากขึ้น แต่ในขณะเดียวกัน ย่อมส่งผลให้สูญเสียทรัพยากรในการทำงานเพิ่มมากขึ้นเช่นกัน ทั้งเวลาในการเก็บข้อมูลที่นานขึ้น ค่าตอบแทนที่มากขึ้น ดังนั้นการหาจุดสมดุลที่ดีจึงสำคัญเป็นอย่างมาก แต่กลับยังไม่ถูกศึกษาวิจัยมากนัก

จากรูปภาพที่ 9 กระบวนการตรวจสอบคุณภาพของข้อมูลบนระบบคราวด์ซอร์สซิงด้วยการเพิ่มกระบวนการในการเก็บข้อมูล ภาพรวมกระบวนการตรวจสอบคุณภาพของข้อมูลทั้งหมดของการเก็บข้อมูล แบ่งออกเป็น 3 ส่วน ได้แก่ 1. งานที่จำเป็นต้องทำก่อน (Job Prerequisite) เป็นงานคัด

กรองผู้ปฏิบัติงานก่อนจะได้ทำงานหลัก ต้องผ่านเกณฑ์ที่กำหนดจึงจะผ่านไปสู่ขั้นตอนถัดไปได้ 2. การแทรกคำถามมาตรฐานแบบทองคำ (Gold Standard Question) เข้าไปในงานหลัก เพื่อตรวจสอบคุณภาพของข้อมูลระหว่างการทำงานของผู้ปฏิบัติงาน หากไม่ผ่านเกณฑ์ที่กำหนด ข้อมูลส่วนนั้นจะถูกคัดออก 3. การทำซ้ำของข้อมูล (Data Redundancy) เมื่อได้ข้อมูลจากงานหลักที่ผ่านเกณฑ์มาแล้ว จะนำข้อมูลที่ label ข้อมูลชุดเดียวกัน มาดูผล majority vote เพื่อสรุปเป็นผลลัพธ์สุดท้าย ให้ได้ข้อมูลที่มีคุณภาพสูงออกมา

งานวิจัยส่วนนี้จึงตั้งใจที่จะทำการทดลอง เก็บผล และวิเคราะห์เกี่ยวกับปริมาณ ชนิดของรูปแบบคำถาม และการนำมาปรับใช้ในการเก็บข้อมูลหนึ่ง ๆ ผ่านคราวด์ซอร์สซิงของไทยชื่อ ว่าง: ตลาดข้อมูล (Wang: Data Market) เพื่อดูผลกระทบที่เปลี่ยนแปลงไปต่อประสิทธิภาพในการคัดกรองคุณภาพของข้อมูล เพื่อที่ผลการวิเคราะห์เหล่านี้จะสามารถใช้เป็นแนวทางให้ Data Requester สำหรับนำไปประยุกต์ใช้ในการออกแบบการเก็บข้อมูลในงานต่าง ๆ ของตัวเองต่อไป

โดยงานวิจัยส่วนนี้ จะทำการทดลองการเพิ่มกระบวนการใบการเก็บข้อมูล 3 แบบ ได้แก่ 1. งานที่จำเป็นต้องทำก่อน (Job Prerequisite) 2. คำถามมาตรฐานทองคำ (Gold Standard Question) และ 3. การทำซ้ำของข้อมูล (Data Redundancy)



รูปภาพที่ 9 กระบวนการตรวจสอบคุณภาพของข้อมูลบนระบบคราวด์ซอร์สซิงด้วยการเพิ่มกระบวนการในการเก็บข้อมูล

4.2.1. งานที่จำเป็นต้องทำก่อน (Job Prerequisite)

การเพิ่มงานที่จำเป็นต้องทำก่อนเข้าไปในกระบวนการเก็บข้อมูล เป็นหนึ่งในกรรมวิธีการคัดกรองผู้ปฏิบัติงานที่ไม่เหมาะสมออกไป กระบวนการคัดกรองนี้สามารถยืนยันได้ว่า ผู้ปฏิบัติงานที่ถูกเลือกให้ผ่านนั้นมีคุณสมบัติ ทักษะ และความสามารถตรงตามความต้องการของผู้ร้องขอข้อมูล และเป็นผู้ปฏิบัติงานที่ได้อ่านและทำความเข้าใจเกี่ยวกับรายละเอียดการทำงานนั้นเรียบร้อยแล้ว

สิ่งที่ผู้ร้องขอข้อมูลมักคำนึงถึงเวลาออกแบบงานที่จำเป็นต้องทำก่อน เช่น ควรจะมีคำถามในงานที่จำเป็นต้องทำก่อนกี่ข้อ ควรออกแบบคำถามในงานที่จำเป็นต้องทำก่อนอย่างไร การทดลองนี้จึงได้ออกแบบคำถามมาทั้งหมด 3 ประเภท ได้แก่

ประเภทที่ 1 คำถามสามัญสำนัก (Common Sense Questions) เป็นคำถามเกี่ยวกับความรู้ทั่วไป โดยไม่จำเป็นต้องเกี่ยวข้องกับงาน ตัวอย่างเช่น มีคลิปเสียงแมวให้ฟัง แล้วให้ระบุประเภทของสัตว์ดังกล่าว หรือถามว่าพระอาทิตย์ขึ้นทางทิศใด

ประเภทที่ 2 คำถามเกี่ยวกับการดำเนินงาน (Logistical Questions) เป็นคำถามเกี่ยวกับขั้นตอนการทำงาน ซึ่งอาจเป็นข้อมูลที่เห็นได้อย่างชัดเจน เช่น ชื่องาน ค่าตอบแทนของงาน หรืออาจเป็นข้อมูลของการทำงานในภาพรวม ที่งานต่าง ๆ ในระบบมีเหมือนกัน เช่น บทลงโทษหากทำงานผิดเงื่อนไขหรือโกง

ประเภทที่ 3 คำถามเกี่ยวกับรายละเอียดงาน (Job Details Questions) เป็นคำถามเกี่ยวกับรายละเอียดงาน เพื่อใช้ในการตรวจสอบ การอ่านรายละเอียดของงานอย่างทั่วถึงของผู้ปฏิบัติงาน และทดสอบความเข้าใจในกระบวนการทำงาน หรือเกณฑ์ในการติดป้ายกำกับข้อมูล

การสร้างงานที่จำเป็นต้องทำก่อน เป็นวิธีหนึ่งที่สามารถคัดกรองผู้ปฏิบัติงานที่ไม่มีคุณภาพออกไปได้ โดยวิธีนี้เป็นการคัดกรองในขั้นตอนก่อนเริ่มทำงาน เพื่อตรวจสอบว่าผู้ปฏิบัติงานที่จะเข้ามาทำงานมีทักษะความสามารถเพียงพอหรือไม่ มีความรู้ความเข้าใจและกรอบความคิด ตรงกันกับสิ่งที่ผู้ร้องขอข้อมูลต้องการหรือเปล่า อ่านรายละเอียดการทำงานครบถ้วนมาแล้วเป็นอย่างดีหรือไม่ เพื่อให้ได้คำตอบหรือคุณภาพของงานที่ตรงกับความต้องการ คนที่ไม่ผ่านเกณฑ์ที่กำหนดจะถูกคัดกรองออกไปให้ไม่สามารถทำงานนี้ได้ นับเป็นการคัดกรองเบื้องต้นอย่างหนึ่ง

การทดลอง

ในการทดลองนี้ งานที่จำเป็นต้องทำก่อน ประกอบด้วย 6 คำถาม ประเภทละ 2 ข้อ แต่ละข้อเป็นคำถามแบบเลือกตอบ ดังแสดงในรูปภาพที่ 10 ผู้ปฏิบัติงานจะผ่านเกณฑ์ก็ต่อเมื่อ ตอบคำถามในงานที่ต้องทำก่อนถูกต้องทั้งหมด 6 ข้อ โดยที่งานจะถูกเปิดในคราวด์ซอร์สซิงแพลตฟอร์ม ว่าง: ตลาดข้อมูล (Wang: Data Market) ซึ่งงานเปิดพร้อมให้ผู้ปฏิบัติงานในระบบที่เข้ามาเห็น สามารถเข้ามาทำได้จนกว่างานจะได้รับครบตามจำนวนที่กำหนดไว้ โดยผู้ปฏิบัติงานสามารถเข้ามาทำงานได้

ในระบบแบบ “มาก่อนได้ก่อน” (“First come-first-served”) ในการทดลองงานที่ต้องทำก่อนนี้ มีผู้ปฏิบัติงานเข้ามาทำจนเสร็จทั้งหมด 2,597 คน และเป็นผู้ผ่านเกณฑ์ 1,291 คน จากนั้นวิเคราะห์ประสิทธิภาพประเภทของคำถาม และดูว่าคำถามเหล่านี้สามารถช่วยคัดกรองผู้ปฏิบัติงานเพิ่มเติมได้อย่างไร

โดยจะเก็บข้อมูลเพื่อนำมาวิเคราะห์ปัจจัยในการออกแบบงานที่ต้องทำก่อนว่า เมื่อมีคำถามเยอะขึ้น จะส่งผลต่อปริมาณคนที่ถูกคัดกรองออกไปอย่างไรบ้าง และคำถามแต่ละประเภท มีศักยภาพในการคัดกรองคนออกแตกต่างกันหรือไม่ อย่างไร และคำถามประเภทใดสามารถคัดกรองคนได้ดีมากที่สุด

การทดลองนี้มีคำถามดังต่อไปนี้

Q1. ภารกิจนี้จะให้ทำแบบทดสอบเรื่องอะไร

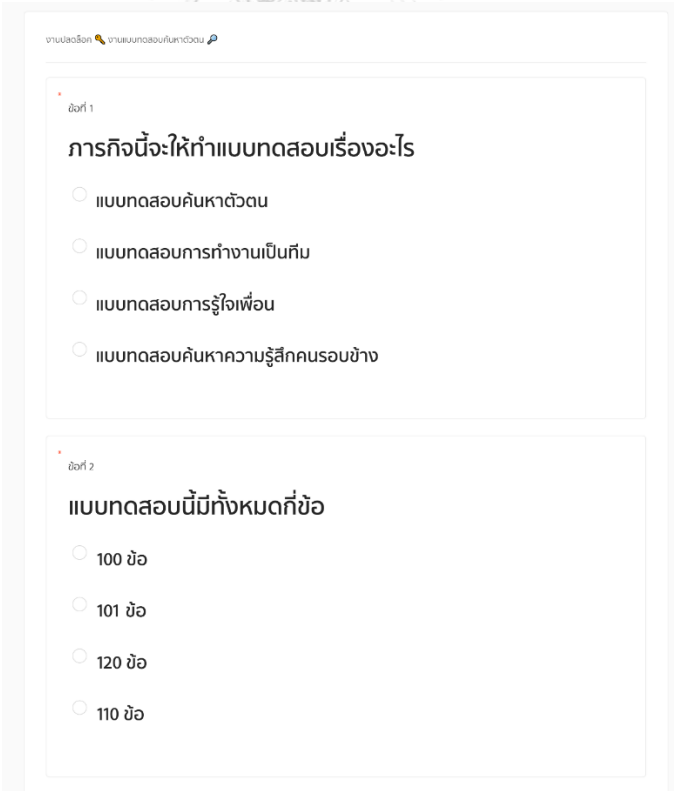
Q2. แบบทดสอบนี้มีทั้งหมดกี่ข้อ

Q3. เสียงที่ได้ยินเป็นเสียงของอะไร

Q4. ข้อใดคือข้อควรระวังในการทำงาน

Q5. โดยปกติแล้ว พระอาทิตย์ขึ้นทางทิศใด

Q6. เมื่อทำการตรวจสอบงาน และพบว่ามีการทำงานที่ไม่เป็นไปตามเงื่อนไข จะเกิดอะไรขึ้น



รูปภาพที่ 10 ภาพแสดงตัวอย่างส่วนต่อประสานผู้ใช้ของคำถามจากงานที่ต้องทำก่อน ประกอบด้วยคำถามแบบเลือกตอบเกี่ยวกับงานที่ต้องทำ

ผลกระทบของปริมาณคำถามต่อปริมาณของคนทำงานที่ถูกคัดกรองออกไป

จำนวนคำถาม	จำนวนเฉลี่ยผู้ปฏิบัติงานที่ผ่านเกณฑ์	จำนวนการเปลี่ยนแปลงเฉลี่ยผู้ปฏิบัติงานที่ผ่านเกณฑ์
1	2246.67	-
2	1969.13	-277.53
3	1746.30	-222.83
4	1564.87	-181.43
5	1415.33	-149.53
6	1291.00	-124.33

ตารางที่ 3 ตารางแสดงจำนวนผู้ปฏิบัติงานที่ผ่านเกณฑ์งานที่ต้องทำก่อน ในกรณีการใช้จำนวนคำถามที่แตกต่างกัน

ผลลัพธ์

อันดับแรก พิจารณาจำนวนคำถามส่งผลต่อจำนวนพนักงานที่จะผ่านการคัดกรองนี้อย่างไร สิ่งนี้ทำได้โดยการจำลองชุดคำถามย่อยต่าง ๆ ที่แตกต่างกัน และสังเกตจำนวนผู้ปฏิบัติงานที่จะผ่านไปตามข้อมูลที่รวบรวมได้ ผลลัพธ์สรุปไว้ในตารางที่ 3

ไม่น่าแปลกใจที่การเพิ่มคำถามที่ต้องทำก่อน (Prerequisite Questions) ส่งผลให้มีจำนวนผู้ปฏิบัติงานที่ถูกคัดกรองออกมากขึ้น โดยเฉลี่ยการเพิ่มคำถามที่ต้องทำก่อนหนึ่งข้อสามารถคัดกรองผู้ปฏิบัติงานเพิ่มอีก 191 คน อย่างไรก็ตาม การเพิ่มจำนวนคำถามในแบบทดสอบยังเพิ่มค่าใช้จ่ายและอาจทำให้ผู้ปฏิบัติงานรู้สึกหนักใจและรู้สึกไม่อยากทำงานได้อีกด้วย ให้สังเกตที่ว่าการเพิ่มความสามารถในการคัดกรองของแต่ละคำถามจะลดลง ดังนั้น ความต้องการพื้นฐานของงานควรมีคำถามเพียงพอที่จะคัดกรองผู้ปฏิบัติงานที่ไม่ตั้งใจทำงานหรือไม่ตรงตามข้อกำหนด พร้อมกับครอบคลุมรายละเอียดสำคัญของงานเพื่อให้แน่ใจว่าความเข้าใจของผู้ปฏิบัติงานตรงกับข้อกำหนด

ประเภท	คำถาม	%ไม่ผ่าน	%ไม่ผ่านเฉพาะคำถามนี้	%ไม่ผ่าน 1 คำถามในประเภท	%ไม่ผ่าน ≥1คำถามในประเภท	%ไม่ผ่าน 1 คำถามในประเภทนี้เท่านั้น	%ไม่ผ่าน ≥1คำถามในประเภทนี้เท่านั้น
1	Q3	10.94	0.15	17.21	17.98	7.32	7.70
	Q5	7.82	8.59				
2	Q1	1.31	4.62	13.71	14.25	3.16	3.16
	Q6	13.48	9.66				
3	Q2	23.49	2.70	29.38	38.39	18.25	22.45
	Q4	23.91	3.00				

ตารางที่ 4 ตารางแสดงสถิติของผู้ปฏิบัติงานที่ผ่านหรือไม่ผ่านในคำถามที่ต้องทำก่อนแต่ละข้อ

ต่อไป จะพิจารณาว่าประเภทของคำถามที่ต้องทำก่อนมีผลอย่างไรต่อความสามารถในการคัดกรอง ผลลัพธ์เกี่ยวกับประสิทธิภาพของคำถามที่ต้องทำก่อนถูกแสดงไว้ในตารางที่ 4 โดยเฉลี่ยแล้ว คำถามเกี่ยวกับรายละเอียดงาน (ประเภทที่ 3) มีประสิทธิภาพสูงสุดในเชิงประสิทธิภาพการคัดกรอง ในขณะที่คำถามเกี่ยวกับการดำเนินงาน (ประเภทที่ 2) มีประสิทธิภาพน้อยที่สุด อย่างไรก็ตาม มีเพียงประมาณครึ่งหนึ่งของผู้ปฏิบัติงาน (1,291 คนจาก 2,597 คน) เท่านั้นที่ผ่านการคัดกรองด้วยคำถามที่ต้องทำก่อนทั้งหมด ซึ่งหมายความว่าคำถามเหล่านี้มีความสัมพันธ์และเป็นส่วนเสริมซึ่งกันและกัน เนื่องจากคำถามทั้งสามประเภทนั้นทดสอบในแง่มุมที่แตกต่างกัน ดังนั้นจึงเป็นการดีกว่าที่จะผสมคำถามทั้งสามประเภทเข้าด้วยกันโดยมีคำถามประเภทที่ 3 เป็นหลัก

ผลของการเพิ่มคำถามในแต่ละประเภทเข้าไปในงานที่ต้องทำก่อนยังถูกวิเคราะห์ด้วย นอกจากคำถามประเภทที่ 3 การเพิ่มคำถามมากกว่าหนึ่งคำถามในแต่ละประเภทมีผลเพียงเล็กน้อยต่อการคัดกรองเท่านั้น ดังนั้นอาจดีกว่าที่จะเน้นคำถามประเภทที่ 3

4.2.2. คำถามมาตรฐานทองคำ (Gold Standard Question)

คำถามมาตรฐานทองคำ คือคำถามที่มีคำตอบที่แน่นอนซึ่งถูกรวมแทรกเข้าไปในชุดคำถามของงานหลัก ถูกสร้างขึ้นเพื่อช่วยตรวจสอบคุณภาพของข้อมูลและความตั้งใจระหว่างการทำงานของผู้ปฏิบัติงาน การมีจำนวนคำถามมาตรฐานทองคำมากเกินไปในงานอาจส่งผลให้ประสิทธิภาพการทำงานลดลง ในขณะที่การมีจำนวนคำถามมาตรฐานทองคำน้อยเกินไปอาจไม่ได้มีผลช่วยตรวจสอบคุณภาพ งานวิจัยนี้ได้ตรวจสอบคำถามมาตรฐานทองคำ 2 ประเภท ได้แก่

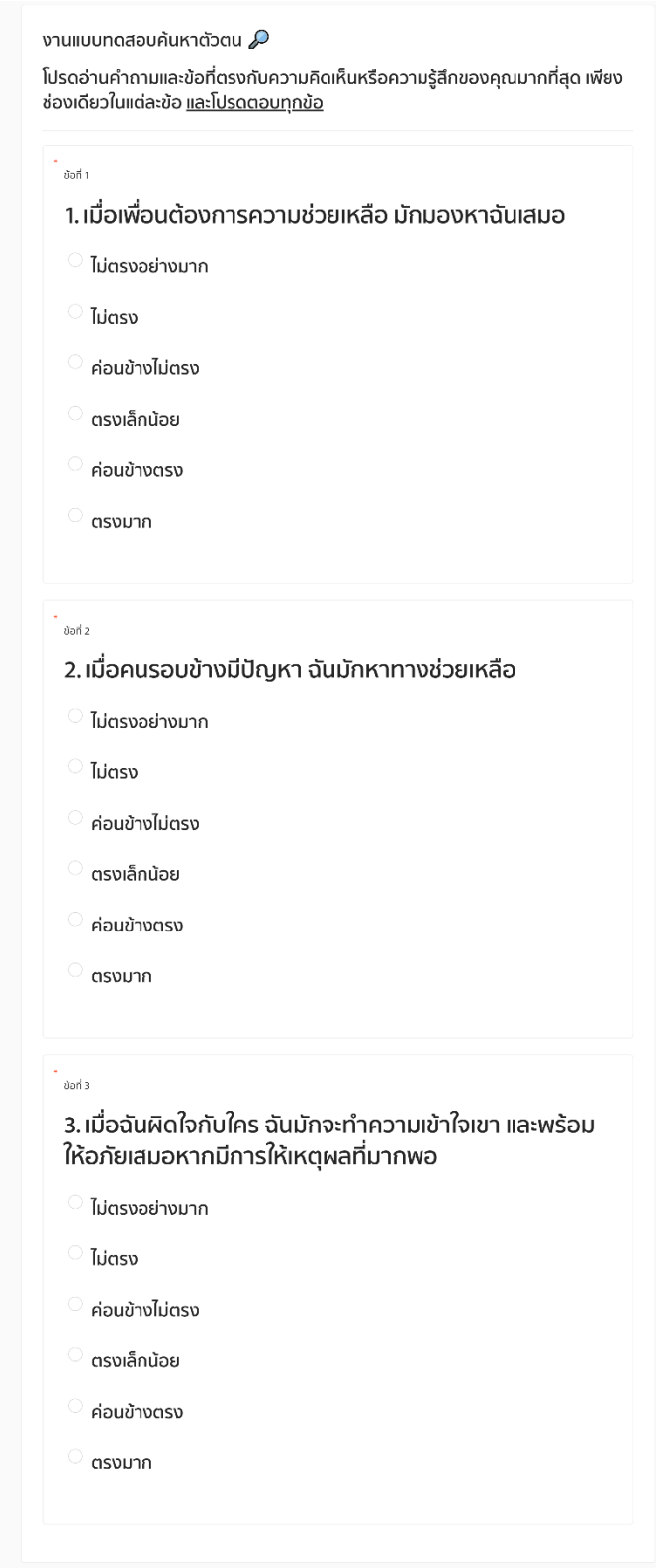
ประเภท 1: คำถามที่ชัดเจน (Obvious Question) เป็นคำถามที่มีคำสั่งหรือคำแนะนำที่" แน่นอน หรือเป็นสิ่งที่รู้จักโดยทั่วกัน เช่น “โปรดตอบคำถามนี้ว่า เห็นด้วยมาก” และ “แมวไม่เห่า” คำถามที่เรียบง่ายนี้สามารถตรวจสอบว่าผู้ปฏิบัติงานอ่านคำถามหรือไม่ และมีประสิทธิภาพในการคัดกรองผู้ปฏิบัติงานที่มีจุดประสงค์ร้ายบางรายได้

ประเภท 2: คำถามที่ซ้ำกัน (Duplicate Question) บางคำถามสามารถใช้ซ้ำหรือเปลี่ยนแปลงเล็กน้อยแล้วถามซ้ำสองครั้งในชุดคำถามเดียวกัน หากคำตอบเหมือนกันถือว่าถูกต้อง คำถามประเภทนี้เหมาะสำหรับการตรวจสอบคำถามที่คำตอบขึ้นกับตัวบุคคล (Subjective) เรื่องเกี่ยวกับความรู้สึกหรือจิตวิทยามากขึ้นซึ่งอาจไม่มีคำตอบแน่นอน อย่างไรก็ตาม ยังสามารถใช้ตรวจสอบความสม่ำเสมอในการทำงานของผู้ใช้ได้

การทดลอง

การทดลองนี้ดำเนินการบนงานแบบสอบถามบุคลิกภาพที่มีข้อความเช่น “คุณกลัวที่จะพูดคุยกับคนแปลกหน้า” แบบสอบถามประกอบด้วยคำถามทั้งหมด 108 ข้อ ไม่รวมคำถามมาตรฐานทองคำ แต่ละคำถามสามารถตอบได้ระดับ 6 ระดับ ได้แก่: ไม่เห็นด้วยอย่างยิ่ง, ไม่เห็นด้วย, ค่อนข้างไม่เห็นด้วย, ค่อนข้างเห็นด้วย, เห็นด้วย, เห็นด้วยอย่างยิ่ง ตัวอย่างส่วนต่อประสานผู้ใช้ของการทดลองนี้แสดงในรูปภาพที่ 11 ชุดคำถามนี้เป็นเครื่องมือประเมินจิตวิทยาหรือแบบทดสอบบุคลิกภาพเพื่อจัดบุคคลออกเป็นกลุ่มต่าง ๆ โดยมีจำนวนผู้ทดสอบเป้าหมายที่ 1,000 คน

คำถามมาตรฐานทองคำถูกเพิ่มในแบบสอบถามดังนี้ คำถามประเภท 1 คำถามที่ชัดเจน 4 ข้อ และคำถามประเภท 2 คำถามที่ซ้ำกัน 4 ข้อ งานจะผ่านเกณฑ์การตรวจสอบคุณภาพเมื่อตอบคำถามทั้ง 8 ข้อของคำถามมาตรฐานทองคำได้อย่างถูกต้อง ผู้ปฏิบัติงานที่สามารถทำงานนี้ได้จำเป็นต้องผ่านการทดสอบจากงานที่ต้องทำก่อนด้วยคำถาม 6 ข้อ ผู้ปฏิบัติงานทั้งหมด 2,597 คน มีผู้ปฏิบัติงานที่ผ่านงานที่ต้องทำก่อนทั้งหมด 1,295 คน และมีจำนวน 1,000 คนได้ถูกเลือกสำหรับการทำแบบสอบถาม โดยเป็นการเลือกแบบ “มาก่อนได้ก่อน”



รูปภาพที่ 11 ภาพแสดงส่วนต่อประสานผู้ใช้งานของงานหลักซึ่งประกอบด้วยคำถามมาตรฐานทองคำที่ถูกแทรกรวมไปกับคำถามทั่วไป เพื่อให้ผู้ปฏิบัติงานไม่ทราบว่าคำถามใดที่ต้องตอบให้ถูกต้อง

อายุ	เพศชาย	เพศหญิง	ไม่ระบุ	รวม
ไม่ได้รับการยืนยัน	8	50	454	512
0 – 18	7	33	0	40
19 – 22	14	81	1	96
23 – 30	28	193	0	221
31 – 45	21	100	0	121
46 – 60	2	7	0	9
61 – 100	1	0	0	1
รวม	81	464	455	1,000

ตารางที่ 5 ตารางแสดงข้อมูลประชากรของผู้ปฏิบัติงาน

ข้อมูลทางสถิติของประชากรผู้ปฏิบัติงานที่ร่วมทดลองถูกแสดงในตารางที่ 5 ข้อมูลอายุได้รับมาจากกระบวนการยืนยันตัวตนของระบบคราวด์ซอร์สซิงแพลตฟอร์มว่าง โดยใช้บัตรประจำตัวประชาชนของประเทศไทย ในขณะที่ข้อมูลเพศถูกรายงานจากตัวผู้ปฏิบัติงานเอง มีผู้ปฏิบัติงานยืนยันตัวตนด้วยบัตรประจำตัวประชาชนเสร็จสิ้นทั้งหมด 488 คน และผู้ปฏิบัติงานที่ให้ข้อมูลเพศจำนวน 545 คน

เพื่อผ่านเกณฑ์ของคำถามมาตรฐานทองคำ มี 2 เงื่อนไข ได้แก่ เงื่อนไขที่ 1 ผู้ปฏิบัติงานต้องตอบคำถามที่ชัดเจน (คำถามประเภทที่ 1) อย่างถูกต้องและเลือกคำตอบเดียวกันสำหรับคู่ของคำถามที่ซ้ำกัน (คำถามประเภทที่ 2) เงื่อนไขที่ 2 ผู้ปฏิบัติงานต้องตอบในทิศทางเดียวกับคำตอบที่ถูกต้องของคำถามที่ชัดเจน และเลือกคำตอบในทิศทางเดียวกันสำหรับคู่ของคำถามที่ซ้ำกัน ทิศทางคำตอบเดียวกันหมายถึงทิศทางที่ไม่เห็นด้วย (ไม่เห็นด้วยอย่างยิ่ง, ไม่เห็นด้วย, ค่อนข้างไม่เห็นด้วย) หรือทิศทางที่เห็นด้วย (ค่อนข้างเห็นด้วย, เห็นด้วย, เห็นด้วยอย่างยิ่ง)

เนื่องจากแบบสอบถามมีคำถามมากกว่า 100 ข้อ จึงเป็นไปได้ยากที่ผู้ปฏิบัติงานจะสามารถให้คำตอบที่เหมือนกันทุกประการในคำถามที่ซ้ำกันได้ ดังนั้นเกณฑ์จึงถูกผ่อนปรนให้ผู้ปฏิบัติงานตอบไปในทิศทางเดียวกันก็นับว่าถูกต้อง ในทำนองเดียวกัน สามารถพิจารณาแบบทิศทางที่ถูกต้องสำหรับคำถามที่ชัดเจนได้เช่นกัน โดยจะใช้คำว่า “แบบแม่นยำ (ม)” (“Exact (E)”) และ "แบบทิศทาง (ท)” (“Directional (D)”) เพื่ออธิบายว่า เกณฑ์ความเข้มงวดใดถูกใช้พิจารณาความถูกต้องของคำตอบ

ผลลัพธ์

จำนวนผู้ปฏิบัติงานที่ตอบคำถามมาตรฐานทองคำได้อย่างถูกต้อง ถูกสรุปไว้ในตารางที่ 6 ผู้ร้องขอข้อมูลถือว่าผู้ปฏิบัติงานที่ผ่านเกณฑ์ คือผู้ที่ตอบถูกต้องในแบบทิศทางสำหรับทั้งคำถามที่ชัดเจนและคำถามที่ซ้ำกัน ส่งผลให้มีผู้ปฏิบัติงานที่ผ่านเกณฑ์ทั้งสิ้น 634 คน

คำถามที่ชัดเจน 4 ข้อ		คำถามที่ซ้ำกัน 4 ข้อ		คำถามที่ชัดเจน 4 ข้อ และคำถามที่ซ้ำกัน 4 ข้อ			
ม	ท	ม	ท	ม ม	ม ท	ท ม	ท ท
739	908	179	693	139	526	163	634

ตารางที่ 6 ตารางแสดงจำนวนผู้ปฏิบัติงานที่ตอบคำถามประเภทต่างๆ ทั้ง 4 ข้อได้อย่างถูกต้องทั้งหมด

จำนวนคำถาม	คำถามที่ชัดเจน	คำถามที่ซ้ำกัน
1	66.80	72.63
2	69.18	80.63
3	71.05	87.73
4	72.60	94.10

ตารางที่ 7 ตารางแสดงเปอร์เซ็นต์ของผู้ปฏิบัติงานที่ไม่ผ่านเกณฑ์การคัดกรองของคำถามมาตรฐานทองคำทั้งสองประเภทในกรณีจำนวนคำถามที่แตกต่างกัน

แม้ว่าการเพิ่มคำถามมาตรฐานทองคำสามารถช่วยกรองการส่งข้อมูลไม่พึงประสงค์ได้ดีขึ้น แต่อาจส่งผลให้เกิดค่าใช้จ่ายที่สูงขึ้นและใช้เวลานานขึ้น การเข้าใจประสิทธิภาพของคำถามทั้ง 2 ประเภทในการคัดกรองผู้ปฏิบัติงานที่ไม่ผ่านเกณฑ์ อาจนำไปสู่กระบวนการคัดกรองที่มีประสิทธิภาพมากขึ้น ด้วยเหตุนี้ จึงจำลองจำนวนคำถามที่แตกต่างกันจากข้อมูลที่รวบรวมมา และพิจารณาผลกระทบต่อประสิทธิภาพการคัดกรองดังตารางที่ 7

คำถามที่ซ้ำกันมีประสิทธิภาพการคัดกรองมากกว่าคำถามที่ชัดเจนอย่างสังเกตได้ การเพิ่มคำถามที่ซ้ำกันแต่ละข้อ ส่งผลให้ประสิทธิภาพการคัดกรองเพิ่มขึ้นประมาณ 7% โดยมีจำนวนข้อมูลของผู้ปฏิบัติงานที่ผ่านเกณฑ์ลดลงเมื่อมีจำนวนคำถามมาตรฐานทองคำเพิ่มขึ้น ในกรณีที่มีเพียงคำถามมาตรฐานทองคำประเภทคำถามที่ซ้ำกัน 4 คำถาม พบว่า ยังมีข้อมูลของผู้ปฏิบัติงานที่ไม่ผ่านเกณฑ์อยู่ประมาณ 6% เช่นเดียวกับการคัดกรองแบบงานที่จำเป็นต้องทำก่อน การผสมประเภทคำถามกันในการใช้งานจริงมีประโยชน์และเหมาะสมมากกว่า

ตารางที่ 8 แสดงผลของการใช้คำถามที่ชัดเจนร่วมกับคำถามที่ซ้ำกัน 4 คำถาม การเพิ่มคำถามที่ชัดเจน 1 ข้อ จะเพิ่มความสามารถการคัดกรองโดยเฉลี่ยประมาณ 1.93% ขึ้นอยู่กับจำนวนคำถามที่ใช้ การตัดสินใจเลือกแลกเปลี่ยน (trade-off) ระหว่างจำนวนคำถามมาตรฐานทองคำและประสิทธิภาพการคัดกรองสามารถพิจารณาตามสถานการณ์ได้อย่างเหมาะสม

จำนวนคำถามที่ซ้ำกัน	จำนวนคำถามที่ชัดเจน	% การคัดกรอง
4	1	96.13
4	2	97.70
4	3	98.95
4	4	100.00

ตารางที่ 8 ตารางแสดงเปอร์เซ็นต์ของผู้ปฏิบัติงานที่ไม่ผ่านเกณฑ์การคัดกรองโดยคำถามที่ซ้ำกัน 4 คำถาม และคำถามที่ชัดเจนในกรณีจำนวนคำถามที่แตกต่างกัน

การอภิปรายเกี่ยวกับการออกแบบคำถามที่ชัดเจน

คำถามที่ชัดเจนถามเกี่ยวกับความรู้ทั่วไป สิ่งสำคัญคือ คำถามต้องชัดเจนและไม่กำกวม เพื่อให้มีผลการกรองที่เรียบง่ายแต่มีประสิทธิภาพ

คำถามที่ชัดเจน ที่ใช้ในการทดลอง มีดังต่อไปนี้:

1. ในคำถามนี้ ให้กดเลือกคำตอบ “ไม่เห็นด้วย”
2. ดวงอาทิตย์ขึ้นทางทิศตะวันตกเสมอ
3. แมวไม่เหา
4. ในคำถามนี้ ให้กดเลือกคำตอบ "ค่อนข้างเห็นด้วย"

คำตอบ	คำถามที่ชัดเจนที่ 1	คำถามที่ชัดเจนที่ 2	คำถามที่ชัดเจนที่ 3	คำถามที่ชัดเจนที่ 4
ไม่เห็นด้วยอย่างยิ่ง	3	796	25	2
ไม่เห็นด้วย	968	148	3	1
ค่อนข้างไม่เห็นด้วย	8	13	9	9
ค่อนข้างเห็นด้วย	7	8	32	8
เห็นด้วย	8	14	7	976
เห็นด้วยอย่างยิ่ง	6	21	924	4
รวม	1,000	1,000	1,000	1,000

ตารางที่ 9 ตารางแสดงจำนวนคำตอบของคำถามที่ชัดเจน

ตารางที่ 9 สรุปผลการตอบคำถาม น่าแปลกใจที่คำถามง่าย ๆ เช่น “ดวงอาทิตย์ขึ้นทางทิศตะวันตกเสมอ” ได้รับคำตอบ “ไม่เห็นด้วย” จำนวนมาก คำถามดังกล่าวไม่เหมาะสมที่จะใช้เป็นคำถามที่ชัดเจน เนื่องจากคำตอบที่ถูกต้องสามารถเป็นได้ทั้ง “ไม่เห็นด้วยอย่างยิ่ง” หรือ “ไม่เห็นด้วย” ซึ่งส่งผลให้ได้รับคำตอบที่ไม่สอดคล้องกัน

คำถามที่ชัดเจนที่ดีควรมีคำตอบที่เป็นไปได้เพียงคำตอบเดียวซึ่งผู้ปฏิบัติงานสามารถตอบได้ทันที เช่น “ในคำถามนี้ ให้กดเลือกคำตอบ “ไม่เห็นด้วย”” ผู้ปฏิบัติงานที่อ่านคำถามควรสามารถตอบคำถามได้อย่างถูกต้อง แนะนำอย่างยิ่งให้ทำการทดสอบคำถามที่ชัดเจนก่อนนำมาใช้เพื่อตรวจจับข้อผิดพลาดที่อาจเกิดขึ้น นอกจากนี้ สิ่งสำคัญคือการวิเคราะห์คำตอบของคำถามที่ชัดเจนเพื่อดูว่าคำถามนั้นอาจออกแบบไม่ดีหรือไม่ อย่างไรก็ตาม การแก้ไขบางอย่าง สามารถนำมาปรับใช้ใน

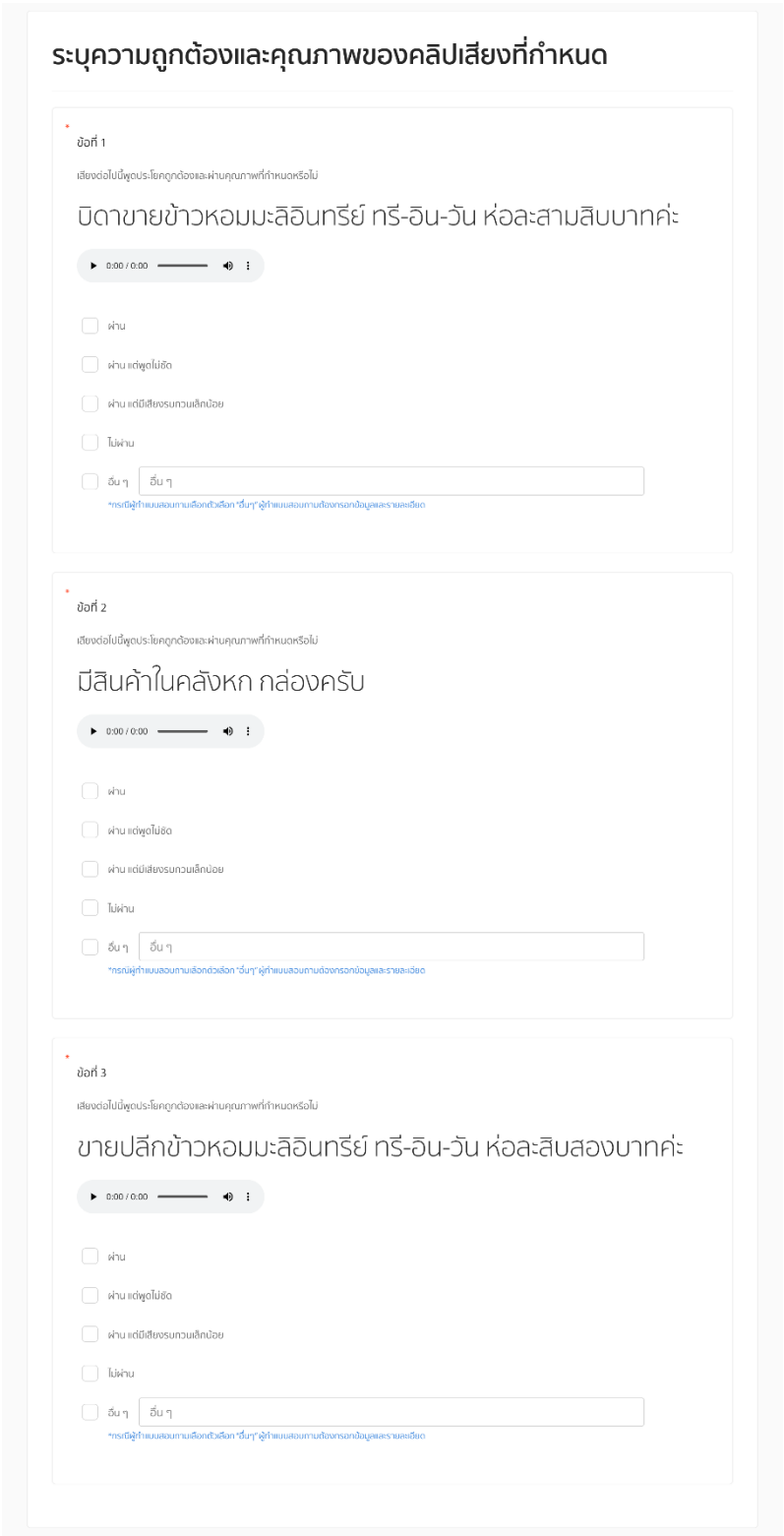
กระบวนการคัดกรองได้ เช่น การใช้เกณฑ์ความเข้มงวดในการตรวจคำตอบแบบทิศทาง หรือการยกเลิกการใช้คำถามที่กำกวมมาเป็นเกณฑ์

4.2.3. การทำซ้ำของข้อมูล (Data Redundancy)

การทำซ้ำของข้อมูลเป็นอีกวิธีหนึ่งในการปรับปรุงคุณภาพข้อมูลในระหว่างกระบวนการติดป้ายกำกับข้อมูล ข้อมูลหนึ่งรายการสามารถถูกติดป้ายโดยผู้ปฏิบัติงานได้มากกว่าหนึ่งคน ถ้าข้อมูลถูกติดป้ายกำกับเหมือนกันจากผู้ปฏิบัติงานทั้ง 2 คน ก็มีความเป็นไปได้สูงที่ป้ายกำกับนั้นจะเป็นคำตอบที่ถูกต้อง อย่างไรก็ตาม คำถามที่สำคัญที่สุดข้อหนึ่งคือ “ข้อมูลจำเป็นต้องตรวจสอบด้วยการทำซ้ำของข้อมูลหรือไม่? ต้องทำซ้ำกี่ครั้ง?" ปริมาณของการทำซ้ำเป็นการแลกเปลี่ยน (trade-off) ระหว่างคุณภาพและทรัพยากรทั้งเงินและเวลา

การทดลอง

สำหรับงานที่ใช้ในการทดลองนี้ เป็นงานตรวจสอบคุณภาพของคลิปเสียงพูด เพื่อนำเสียงไปใช้ในการสร้างระบบรู้จำเสียงพูด (Auto Speech Recognition: ASR) ผู้ปฏิบัติงานต้องฟังคลิปเสียงเพื่อตรวจสอบความถูกต้องของคำพูดในคลิปเสียงเทียบกับประโยคที่กำหนดให้ และตรวจสอบคุณภาพของคลิปเสียง โดยจำแนกคลิปเสียงเป็น 4 ประเภท ได้แก่ 1. ผ่าน 2. ผ่าน แต่พูดไม่ชัด 3. ผ่าน แต่มีเสียงรบกวนเล็กน้อย และ 4. ไม่ผ่าน ดังแสดงตัวอย่างงานในรูปภาพที่ 12 หนึ่งงานประกอบด้วยคลิปเสียงจำนวน 21 คลิป โดยมีคลิปเสียงที่ต้องติดป้ายกำกับทั้งหมด 41,643 คลิป แต่ละคลิปเสียงจะถูกติดป้ายกำกับโดยผู้ปฏิบัติงาน 3 คน การลงคะแนนเสียงส่วนใหญ่ (Majority Vote) จะถูกใช้เป็นตัวบ่งชี้คุณภาพจริงหรือป้ายกำกับผลลัพธ์ของคลิปนั้น จากนั้นข้อมูลการลงคะแนนนี้จะถูกใช้ในการจำลองสถานการณ์ที่แตกต่างกันและตรวจสอบผลกระทบที่เกิดขึ้นต่อการควบคุมคุณภาพ วิธีการเลือกกลุ่มตัวอย่างของผู้ปฏิบัติงานคือ "มาก่อนได้ก่อน" เช่นเดียวกับวิธีการเลือกกลุ่มตัวอย่างผู้ปฏิบัติงานในการทดลองของงานที่ต้องทำก่อน



รูปภาพที่ 12 ภาพแสดงส่วนต่อประสานผู้ใช้ของงานในการทดลองการทำซ้ำของข้อมูล ผู้ปฏิบัติงานถูกขอให้ฟังคลิปเสียงและอ่านประโยคที่กำหนด เพื่อตรวจสอบความถูกต้องและคุณภาพของคลิปเสียง

การทดลองนี้พิจารณาทั้งหมด 5 สถานการณ์ (Scenario: S) ดังต่อไปนี้:

1. สถานการณ์ที่ 1 (S1): ไม่มีการคัดกรอง ทุกเสียงถูกตรวจว่าผ่าน
2. สถานการณ์ที่ 2 (S2): ผู้ปฏิบัติงาน 1 คน ซึ่งเป็นผู้ปฏิบัติงานรายแรกที่ตรวจ
3. สถานการณ์ที่ 3 (S3): ผู้ปฏิบัติงาน 1 คน ซึ่งถูกเลือกโดยการสุ่มจากผู้ปฏิบัติงาน 3 คน
4. สถานการณ์ที่ 4 (S4): ผู้ปฏิบัติงาน 2 คน ถ้าหนึ่งในผู้ปฏิบัติงานตัดสินคุณภาพของคลิปเสียงว่า “ไม่ผ่าน” จะถือว่าไม่ผ่าน
5. สถานการณ์ที่ 5 (S5): ผู้ปฏิบัติงาน 2 คน หากทั้งสองมีความเห็นไม่ตรงกัน ความเห็นของผู้ปฏิบัติงานคนที่ 3 จะถูกนำมาใช้ตัดสิน

	tp	fp	tn	fn	precision	recall	F1
S1	0	0	28869	12774	NaN	0.0000	NaN
S2	12333	2030	26839	441	0.8587	0.9655	0.9089
S3	12228	1218	27651	546	0.9094	0.9573	0.9327
S4	12774	2522	26347	0	0.8351	1.0000	0.9102
S5	12774	0	28869	0	1.0000	1.0000	1.0000

ตารางที่ 10 ตารางแสดงจำนวนของ ค่าจริง (True Positives: tp), ค่าเท็จบวก (False Positives: fp), ค่าจริงลบ (True Negatives: tn), ค่าเท็จลบ (False Negatives: fn), ค่าความเที่ยง (Precision), ค่าความระลึกได้ (recall), และ ค่าคะแนนเอฟวัน (f1) สำหรับแต่ละสถานการณ์ ผลลัพธ์ของป้ายกำกับ “ไม่ผ่าน” เป็นประเภทบวก (Positive Class) ในขณะที่ส่วนที่เหลือถือว่าเป็นประเภทลบ (Negative Class)

ผลลัพธ์

จากตารางที่ 10 สามารถสรุปผลลัพธ์ได้โดยพิจารณางานนี้เป็นงานตรวจจับ ซึ่งจะพิจารณาผลป้ายกำกับ “ไม่ผ่าน” เป็นประเภทบวก (Positive Class) และส่วนที่เหลือถือเป็นประเภทลบ (Negative Class) สถานการณ์ที่ 1 (ไม่มีการคัดกรอง) มีความแม่นยำต่ำที่สุดที่ 69.33% ซึ่งไม่เพียงพอที่จะนำข้อมูลนี้ไปใช้ได้ สถานการณ์ที่ 2 และ สถานการณ์ที่ 3 มีความแม่นยำใกล้เคียงกันที่ 94.07% และ 95.76% ตามลำดับ ซึ่งสะท้อนถึงข้อสำคัญว่า ไม่มีนัยสำคัญในลำดับของผู้ปฏิบัติงาน สิ่งนี้ทำหน้าที่แสดงถึงการตรวจสอบสถานการณ์จำลองของการทดลองนี้ ว่าสามารถใช้งานได้ สถานการณ์ที่ 4 มีความแม่นยำที่ 93.94% ซึ่งต่ำกว่าสถานการณ์ที่ 2 และ สถานการณ์ที่ 3 อย่างไรก็ตาม มีการเรียกคืนที่สมบูรณ์แบบสำหรับเสียงที่ไม่ดี นี่เป็นผลกระทบโดยตรงจากวิธีการเลือกเงื่อนไข

แบบกรณีที่แย่ที่สุดในการตัดสินใจ สถานการณ์ที่ 5 มีผลลัพธ์เหมือนกับการลงคะแนนเสียงส่วนใหญ่ แต่มีเพียง 9% ของคลิปเท่านั้นที่ต้องการผู้ปฏิบัติงานคนที่สาม ซึ่งช่วยลดความต้องการในการใช้แรงงานลงอย่างมาก

ในเชิงของการแลกเปลี่ยน (Trade-off) การใช้ผู้ปฏิบัติงานเพียงคนเดียวก็ให้ประสิทธิภาพในการคัดกรองสูง สามารถคัดกรองข้อมูลที่ไม่ผ่านส่วนใหญ่ได้ (ค่าเรียกคืนประมาณ 0.96) การเพิ่มจำนวนผู้ปฏิบัติงานจะทำให้ประสิทธิภาพเพิ่มขึ้นเพียงเล็กน้อยเท่านั้น อย่างไรก็ตาม ในสถานการณ์ที่คุณภาพข้อมูลสำคัญเป็นอย่างมาก อาจจำเป็นต้องใช้จำนวนผู้ปฏิบัติงานเพิ่มมากขึ้น

4.3. การตรวจสอบคุณภาพของข้อมูลบนคราวด์ซอร์สซิงด้วยแบบจำลองการเรียนรู้ของเครื่อง

4.3.1. ข้อมูลพฤติกรรมการปฏิบัติงานที่เก็บได้จากคราวด์ซอร์สซิงแพลตฟอร์ม “ว่าง”

จากการทำงานของผู้ปฏิบัติงานบนคราวด์ซอร์สซิงแพลตฟอร์ม “ว่าง” จะสามารถเก็บข้อมูลพฤติกรรมในการทำงานที่น่าสนใจได้ ดังตารางที่ 11

ตารางที่ 11 ตารางแสดงข้อมูลพฤติกรรมการปฏิบัติงานที่เก็บได้จากคราวด์ซอร์สซิงแพลตฟอร์ม “ว่าง”

ลำดับที่	รายการ
1	จำนวนครั้งที่ผู้ปฏิบัติงานทำงานนั้น
2	จำนวนวันที่ผู้ปฏิบัติงานทำงานนั้น
3	จำนวนครั้งที่ผู้ปฏิบัติงานเปลี่ยนคำตอบก่อนส่ง
4	จำนวนความยาวของคำตอบ (ในกรณีคำตอบเป็นข้อความ)
5	เวลาเริ่มทำงาน
6	เวลาที่ทำงานเสร็จ
7	ระยะเวลาที่ผู้ปฏิบัติงานใช้ในการทำงานทั้งหมด
8	ระยะเวลาที่ผู้ปฏิบัติงานใช้ในการทำงาน แต่ไม่ได้อยู่ในหน้าการทำงาน (มีการหลุดโฟกัส)
9	ระยะเวลาที่ผู้ปฏิบัติงานใช้ในการทำงานจริง
10	จำนวนลงบันทึก (log) การเคลื่อนที่ของเมาส์ (mouse move)
11	ตำแหน่งของเมาส์ในการเคลื่อนที่แต่ละครั้ง
12	เวลาที่ลงบันทึกตำแหน่งของเมาส์ในการเคลื่อนที่แต่ละครั้ง
13	ระยะเวลาระหว่างการลงบันทึกตำแหน่งของเมาส์ในการเคลื่อนที่แต่ละครั้ง
14	จำนวนครั้งในการคลิกเมาส์ (mouse click)
15	ตำแหน่งของการคลิกเมาส์
16	เวลาที่คลิกเมาส์
17	ระยะเวลาระหว่างการคลิกเมาส์แต่ละครั้ง
18	จำนวนครั้งที่มีการดับเบิ้ลคลิกเมาส์ (double click)
19	ตำแหน่งของการดับเบิ้ลคลิกเมาส์
20	เวลาที่ดับเบิ้ลคลิกเมาส์
21	ระยะเวลาระหว่างการดับเบิ้ลคลิกเมาส์แต่ละครั้ง

ลำดับที่	รายการ
22	จำนวนครั้งในการกดเมาส์ลง (mouse down) ในการทำงาน
23	ตำแหน่งของการกดเมาส์ลง
24	เวลาที่กดเมาส์ลง
25	จำนวนครั้งในการปล่อยเมาส์ขึ้น (mouse up) ในการทำงาน
26	ตำแหน่งของการปล่อยเมาส์ขึ้น
27	เวลาที่ปล่อยเมาส์ขึ้น
28	ระยะเวลาระหว่างการกดเมาส์ลงและการปล่อยเมาส์ขึ้นแต่ละครั้ง
29	จำนวนครั้งในการลงบันทึกการเลื่อน (scroll) เมาส์
30	ตำแหน่งที่ทำการเลื่อนเมาส์
31	เวลาที่ทำการเลื่อนเมาส์
32	ระยะเวลาระหว่างการเลื่อนเมาส์แต่ละครั้ง
33	จำนวนครั้งที่ปุ่มบนคีย์บอร์ดถูกกด (key press)
34	ปุ่มบนคีย์บอร์ดที่ถูกกด
35	เวลาที่ปุ่มคีย์บอร์ดนั้นถูกกด
36	ระยะเวลาระหว่างการกดปุ่มบนคีย์บอร์ดแต่ละครั้ง
37	จำนวนครั้งที่ปุ่มบนคีย์บอร์ดถูกกดลง (key down)
38	ปุ่มบนคีย์บอร์ดที่ถูกกดลง
39	เวลาที่ปุ่มคีย์บอร์ดนั้นถูกกดลง
40	จำนวนครั้งที่ปุ่มบนคีย์บอร์ดถูกปล่อยขึ้น (key up)
41	ปุ่มบนคีย์บอร์ดที่ถูกปล่อยขึ้น
42	เวลาที่ปุ่มคีย์บอร์ดนั้นถูกปล่อยขึ้น
43	ระยะเวลาระหว่างการที่ปุ่มบนคีย์บอร์ดถูกกดลงและปล่อยขึ้นแต่ละครั้ง
44	จำนวนครั้งที่มีการคัดลอก (copy)
45	เวลาที่มีการคัดลอก
46	จำนวนครั้งที่มีการตัด (cut)
47	เวลาที่มีการตัด
48	จำนวนครั้งที่มีการวาง (paste)
49	เวลาที่มีการวาง
50	ระยะเวลาระหว่างการคัดลอก การตัด และการวางแต่ละครั้ง

ลำดับที่	รายการ
51	จำนวนครั้งที่มีการโฟกัสหน้าต่าง
52	เวลาที่โฟกัสหน้าต่าง
53	จำนวนครั้งที่มีการหลุดโฟกัสหน้าต่าง
54	เวลาที่หลุดโฟกัสหน้าต่าง
55	ขนาดหน้าต่าง (window size) เริ่มต้น
56	จำนวนครั้งในการเปลี่ยนขนาดหน้าต่าง
57	ขนาดหน้าต่างที่เปลี่ยนแปลงไปในแต่ละครั้ง
58	เวลาที่ทำการเปลี่ยนขนาดหน้าต่าง
59	ระยะเวลาระหว่างการเปลี่ยนขนาดหน้าต่างในแต่ละครั้ง
60	ตำแหน่งหน้าต่างบนหน้าจอ

4.3.2. การออกแบบขั้นตอนกระบวนการทำงานของระบบจำแนกพฤติกรรม และเลือกขั้นตอนกระบวนการทำงานที่เหมาะสม

การเก็บข้อมูลหรือการติดป้ายกำกับข้อมูลด้วยคราวด์ซอร์สซิงแพลตฟอร์ม สามารถกระจายการเก็บข้อมูลออกเป็นส่วนย่อย ๆ เพื่อให้คนบนโลกออนไลน์สามารถเข้ามาช่วยกันทำได้โดยสะดวก ส่งผลให้สามารถเก็บข้อมูลได้หลากหลายและปริมาณมากขึ้น ภายในระยะเวลาที่น้อยลงได้ แต่ในขณะเดียวกัน ผู้ปฏิบัติงานสามารถเป็นใครก็ได้ ทำงานที่ไหนในโลกก็ได้ ส่งผลให้อาจมีปัญหาในเรื่องคุณภาพของข้อมูลที่ได้รับมา ดังนั้นการตรวจสอบคุณภาพของข้อมูลที่ได้รับจากคราวด์ซอร์สซิงแพลตฟอร์มจึงมีความสำคัญเป็นอย่างยิ่งเพื่อลดปัญหาดังกล่าวลง

การตรวจสอบคุณภาพของข้อมูลบนคราวด์ซอร์สซิงด้วยการเพิ่มกระบวนการในการเก็บข้อมูลทั้ง 3 วิธี อาจจะสามารถช่วยคัดกรองผู้ปฏิบัติงานที่มีความรู้ ความสามารถ และทักษะที่ไม่เป็นไปตามความต้องการออกไปได้ แต่ก็มาพร้อมการแลกเปลี่ยนที่เมื่อเพิ่มกระบวนการและคำถามเข้าไปในงาน จะส่งผลให้ต้องเสียทรัพยากรในการเก็บข้อมูลเพิ่มมากขึ้นทั้งแรงงาน เงิน และระยะเวลา จึงต้องพิจารณาการใช้งานอย่างเหมาะสมกับความต้องการ จึงเป็นที่มาของการศึกษาและทดลองการนำแบบจำลองการเรียนรู้ของเครื่อง (Machine Learning Model) มาใช้ตรวจสอบคุณภาพของข้อมูลบนคราวด์ซอร์สซิง

จากการศึกษางานวิจัยที่เกี่ยวข้อง พบว่า ปกติการใช้แบบจำลองการเรียนรู้ของเครื่องเข้ามาช่วยตรวจสอบคุณภาพของข้อมูล การฝึกสอนแบบจำลอง 1 แบบจำลอง โดยใช้ข้อมูลจากงานหนึ่ง ๆ บนคราวด์ซอร์สซิง เมื่อฝึกสอนแบบจำลองดังกล่าวเสร็จ จะสามารถนำแบบจำลองมาใช้ได้กับงานที่

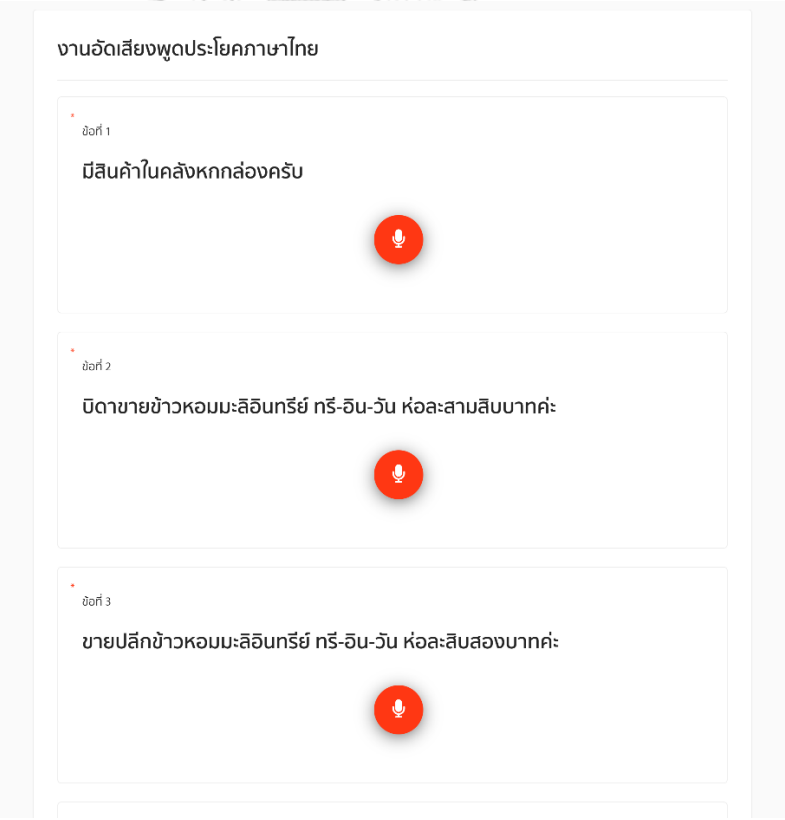
นำข้อมูลมาฝึกสอนแบบจำลองเพียงงานเดียวเท่านั้น งานวิจัยนี้จึงต้องการศึกษาและทดลองการนำกระบวนการฝึกสอนแบบจำลอง และแบบจำลอง ให้สามารถนำมาประยุกต์ใช้ได้กับงานที่หลากหลายบนคราวด์ซอร์สซิงแพลตฟอร์ม

จากงานวิจัยที่เกี่ยวข้อง ได้ทดลองนำข้อมูลพฤติกรรมการทำงานของผู้ปฏิบัติงานมาใช้สร้างแบบจำลองเพื่อตรวจสอบคุณภาพของข้อมูลบนคราวด์ซอร์สซิงแพลตฟอร์ม โดยเมื่อผู้ปฏิบัติงานเข้ามาทำงาน ระบบจะบันทึกข้อมูลลักษณะพฤติกรรมการทำงานของผู้ปฏิบัติงานซึ่งเป็นพฤติกรรมที่ต้องการตรวจสอบเพื่อนำมาใช้ในการฝึกฝนแบบจำลอง และใช้ในการทดลองต่อไป

4.3.3. การสร้างและการฝึกฝนแบบจำลอง

การวิจัยนี้จะทำการทดลองใน 2 งาน ที่เก็บข้อมูลจากคราวด์ซอร์สซิงแพลตฟอร์ม ได้แก่ 1. งานอัดเสียงพูด (Audio Recording) ประโยคภาษาไทย 2. งานตรวจสอบคุณภาพของคลิปเสียงพูด

1. งานอัดเสียงพูด (Audio Recording) ประโยคภาษาไทย เป็นงานที่ให้ผู้ปฏิบัติงานเข้ามาอัดเสียง โดยอัดเสียงอ่านข้อความตามประโยคที่กำหนดให้



รูปภาพที่ 13 ภาพแสดงส่วนต่อประสานผู้ใช้ของงานอัดเสียงพูดประโยคภาษาไทย

2. งานตรวจสอบคุณภาพของคลิปเสียงพูด เป็นงานที่ให้ผู้ปฏิบัติงานเข้ามาฟังคลิปเสียงที่กำหนดให้ และตรวจสอบความถูกต้องของคลิปเสียงโดยเทียบกับประโยคที่กำหนด และตรวจสอบคุณภาพของคลิปเสียง โดยจำแนกคลิปเสียงเป็น 4 ประเภท ได้แก่ 1. ผ่าน 2. ผ่าน แต่พูดไม่ชัด 3. ผ่าน แต่มีเสียงรบกวนเล็กน้อย และ 4. ไม่ผ่าน โดยตอนท้ายที่สุด แต่ละคลิปเสียงจะถูกจัดจำแนกออกเป็น 2 ประเภท ได้แก่ 1. ผ่าน (รวมจาก3 ประเภท ได้แก่ 1. ผ่าน 2. ผ่าน แต่พูดไม่ชัด 3. ผ่าน แต่มีเสียงรบกวนเล็กน้อย) และ 2. ไม่ผ่าน (มาจากประเภทที่ 4. ไม่ผ่าน) ซึ่งอาจมองได้ว่า งานตรวจสอบคุณภาพของคลิปเสียงพูดนี้ เป็นงานประเภท งานจำแนกประเภท (Classification Task) อย่างหนึ่ง

การสร้างและฝึกฝนแบบจำลอง แบ่งออกเป็น 3 ขั้นตอน ดังนี้

1) การฝึกฝนแบบจำลองที่ต้องการพิจารณาด้วยข้อมูลที่เก็บรวบรวมมา

แบบจำลองที่เลือกใช้ในการสร้างแบบจำลองการเรียนรู้ของเครื่องที่ใช้ตรวจสอบคุณภาพของข้อมูลบนคราวด์ซอร์สซิงซึ่งจะนำมาพิจารณาในการทดลองนี้มี 2 แบบจำลอง ได้แก่ 1. แบบจำลองเอ็กซ์จีบูสต์ (XG Boost) หรือ แบบจำลองเอ็กซ์ตรีมเกรเดียนบูสติง (Extreme Gradient Boosting) 2. แบบจำลองเอสวีเอ็ม (SVM) หรือ แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) เนื่องจากทั้งสองแบบจำลองเป็นแบบจำลองขนาดเล็กที่สามารถฝึกสอนได้ง่าย รวดเร็ว และสามารถนำไปประยุกต์ใช้ได้กับงานที่หลากหลายบนคราวด์ซอร์สซิงแพลตฟอร์มได้

โดยข้อมูลนำเข้าของแบบจำลอง เป็นข้อมูลพฤติกรรมการทำงานของผู้ปฏิบัติงาน ที่เก็บรวบรวมมาจากตัวตรวจจับกิจกรรมบนเว็บเบราว์เซอร์ (Browser Event Listener) และระยะเวลาในการทำงานของผู้ปฏิบัติงาน และให้แบบจำลองทำนายผลลัพธ์เพื่อจำแนกว่าพฤติกรรมที่ต้องการตรวจสอบนี้เป็นพฤติกรรมที่ทำให้ข้อมูลผลลัพธ์มีคุณภาพสูงหรือต่ำ โดยตอนฝึกฝนแบบจำลองและวัดผลแบบจำลองจะใช้ข้อมูลการลงคะแนนเสียงส่วนใหญ่ (Majority Vote) จากผู้ปฏิบัติงานที่มาประเมินความถูกต้องและคุณภาพของคลิปเสียงพูด 3 คน ต่อคลิปเสียงพูด เป็นตัวบ่งชี้คุณภาพจริง (Ground Truth) หรือป้ายกำกับผลลัพธ์ของคลิปเสียงนั้น

2) การเลือกแบบจำลองสำหรับแต่ละงาน

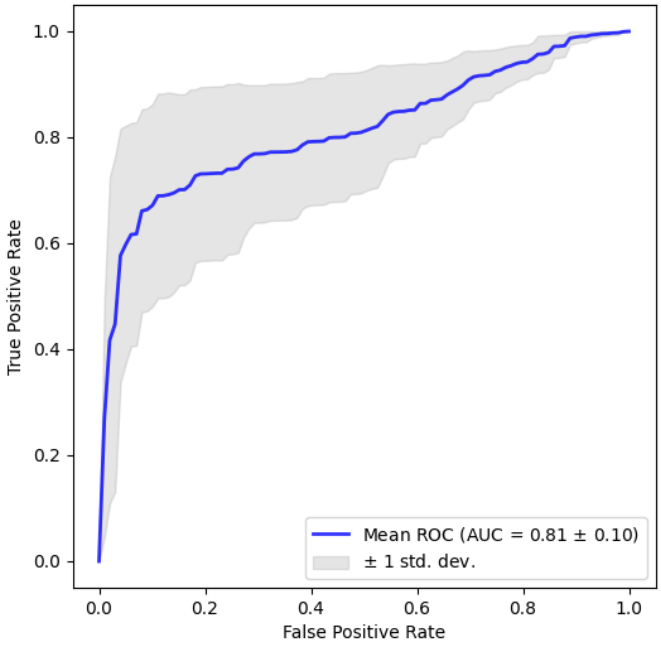
เนื่องจากการตรวจสอบคุณภาพของข้อมูลเป็นการจำแนกประเภทแบบทวิภาค (Binary Classification) กล่าวคือ มีผลการทำนายจำแนกออกเป็น 2 กลุ่ม ได้แก่ 1. ผ่าน หรือ ประเภทลบ (Negative Class) หรือ พฤติกรรมที่ทำให้ข้อมูลผลลัพธ์มีคุณภาพสูง และ 2. ไม่ผ่าน หรือ ประเภทบวก (Positive Class) หรือ พฤติกรรมที่ทำให้ข้อมูลผลลัพธ์มีคุณภาพต่ำ จึงสามารถใช้เส้นโค้งอาร์โอซี หรือเส้นโค้งรีซีฟเวอร์โอเปอเรติงแคแรกเทอริสติก (Receiver Operating Characteristic Curve: ROC curve) เพื่อดูประสิทธิภาพในการทำนายผลการจำแนกประเภทของข้อมูลได้ โดยจะนำข้อมูล

ทั้งหมดมาประมวลผล ด้วยกรรมวิธีการแบ่งข้อมูลเพื่อนำมาทดสอบประสิทธิภาพ เค ส่วน (K-Fold Cross Validation) โดยใช้ค่า K = 5 หรือแบ่งข้อมูลออกเป็น 5 ส่วน จากนั้น วัดผลเปรียบเทียบประสิทธิภาพระหว่างแบบจำลองจากการให้คะแนนด้วยค่าพื้นที่ใต้กราฟ (Area under the ROC Curve: AUC) โดยวัดพื้นที่ใต้กราฟ ถ้าแบบจำลองใดมีพื้นที่ใต้กราฟมากกว่า หมายความว่าแบบจำลองนั้นมีประสิทธิภาพการทำนายเพื่อจำแนกผลลัพธ์ที่ดีกว่า นอกจากนี้ เนื่องจากการประมวลผลโดยมีการแบ่งข้อมูลเพื่อนำมาทดสอบประสิทธิภาพแบบ 5 ส่วน (5 Fold Cross Validation) จึงมีการวัดค่าความแปรปรวน (Variance) เพื่อบ่งบอกการกระจายตัวของข้อมูล และความแปรปรวนของผลลัพธ์ว่ามีเสถียรภาพมากน้อยเพียงใด

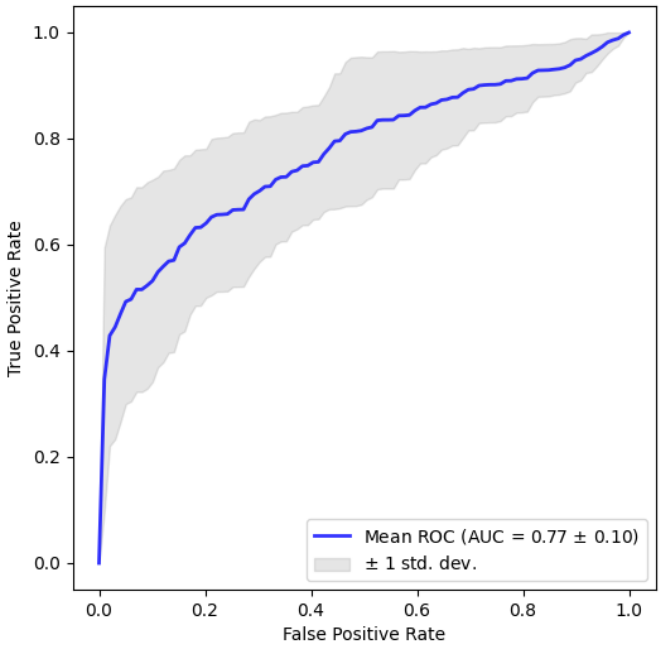
ผลลัพธ์

งาน	แบบจำลองเอสวีเอ็ม		แบบจำลองเอ็กซ์จีบูสต์	
	พื้นที่ใต้กราฟ	ความแปรปรวน	พื้นที่ใต้กราฟ	ความแปรปรวน
งานที่ 1	0.81	0.10	0.77	0.10
งานที่ 2	0.55	0.05	0.88	0.02

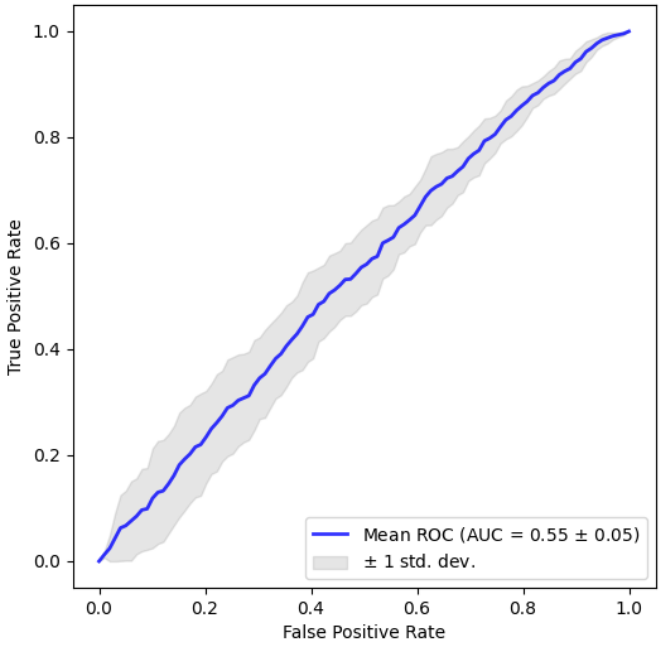
ตารางที่ 12 ตารางแสดงผลค่าพื้นที่ใต้กราฟของเส้นโค้งอาร์โอซีและค่าความแปรปรวนของแบบจำลองในงานประเภทต่าง ๆ



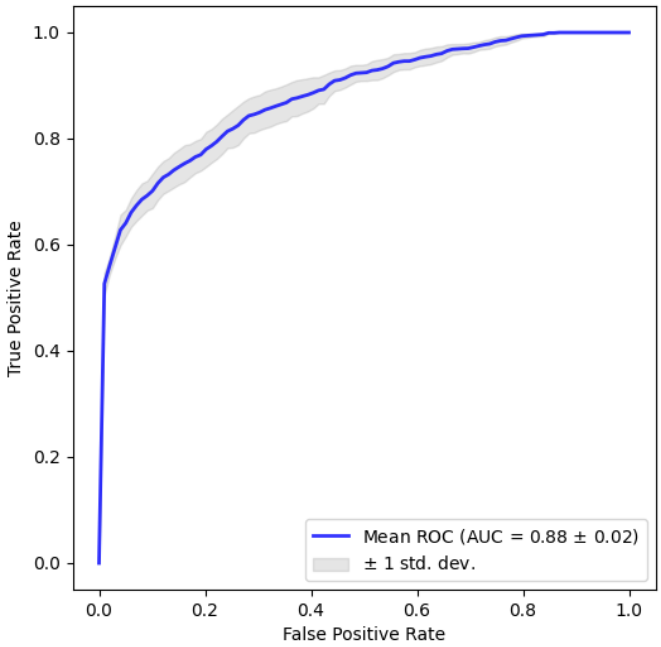
รูปภาพที่ 14 ภาพแสดงเส้นโค้งอาร์โอซีของแบบจำลองเอสวีเอ็มในงานอัดเสียงพูดประโยคภาษาไทย



รูปภาพที่ 15 ภาพแสดงเส้นโค้งอาร์โอซีของแบบจำลองเอ็กซ์จีบูสต์ในงานอัดเสียงพูดประโยคภาษาไทย



รูปภาพที่ 16 ภาพแสดงเส้นโค้งอาร์โอซีของแบบจำลองเอสวีเอ็มในงานตรวจสอบคุณภาพของคลิปเสียงพูด



รูปภาพที่ 17 ภาพแสดงเส้นโค้งอาร์โอซีของแบบจำลองเอ็กซ์จีบูสต์ในงานตรวจสอบคุณภาพของคลิปเสียงพูด

จากผลการทดลองดังแสดงในตารางที่ 12 เห็นได้ว่า งานที่ 1 งานอัดเสียงพูดประโยคภาษาไทย เมื่อใช้แบบจำลองเอสวีเอ็มให้ค่าพื้นที่ใต้กราฟ 0.81 ซึ่งมากกว่าแบบจำลองเอ็กซ์จีบูสต์ที่ให้ค่าพื้นที่ใต้กราฟ 0.77 โดยทั้งสองแบบจำลองมีค่าความแปรปรวน 0.1 ดังนั้นจึงเลือกแบบจำลองเอสวีเอ็มสำหรับใช้ทำการทดลองในงานที่ 1 และงานที่ 2 งานตรวจสอบคุณภาพของคลิปเสียงพูด มีค่าพื้นที่ใต้กราฟ 0.55 ค่าความแปรปรวน 0.05 เมื่อใช้แบบจำลองเอสวีเอ็ม ซึ่งพื้นที่ใต้กราฟน้อยกว่าแบบจำลองเอ็กซ์จีบูสต์ ที่ได้พื้นที่ใต้กราฟ 0.88 ค่าความแปรปรวน 0.02 ดังนั้นจึงเลือกแบบจำลองเอ็กซ์จีบูสต์ใช้ทำการทดลองในงานที่ 2

3) การเลือกขีดแบ่งและนำผลการทำนายไปจำลองในสถานการณ์ที่แตกต่างกัน

เมื่อเลือกแบบจำลองเพื่อใช้ในแต่ละงานแล้ว จึงนำแบบจำลองของแต่ละงานมาใช้ทดลองจำลองสถานการณ์ต่าง ๆ ในการใช้แบบจำลองร่วมกับความเห็นของผู้ปฏิบัติงานในการตรวจสอบคุณภาพของข้อมูล และตรวจสอบผลกระทบที่เกิดขึ้น โดยการทดลองนี้พิจารณาทั้งหมด 5 สถานการณ์ (Scenario: S) ในแต่ละสถานการณ์จะคำนวณค่าผลเฉลี่ยจากทุกวิธีเรียงสับเปลี่ยนและวิธีจัดหมู่ (Permutation and Combination) โดยมี 5 สถานการณ์ ดังต่อไปนี้:

1. สถานการณ์ที่ 1 (S1): ผู้ปฏิบัติงาน 3 คน ตรวจสอบข้อมูล และตัดสินโดยการลงคะแนนเสียงส่วนใหญ่
2. สถานการณ์ที่ 2 (S2): ผู้ปฏิบัติงาน 1 คน ตรวจสอบข้อมูล
3. สถานการณ์ที่ 3 (S3): แบบจำลองทำนายผลการตรวจสอบ หากผลการทำนายอยู่ในช่วงขีดแบ่งที่ไม่มั่นใจ ให้นำความเห็นจากผู้ปฏิบัติงาน 1 คน เข้ามาตรวจสอบข้อมูล โดยเชื่อความเห็นจากผู้ปฏิบัติงาน 1 คนนั้น
4. สถานการณ์ที่ 4 (S4): แบบจำลองทำนายผลการตรวจสอบ หากผลการทำนายอยู่ในช่วงขีดแบ่งที่ไม่มั่นใจ ให้นำความเห็นจากผู้ปฏิบัติงาน 2 คน เข้ามาตรวจสอบข้อมูล โดยตัดสินด้วยการลงคะแนนเสียงส่วนใหญ่ จากแบบจำลองและผู้ปฏิบัติงาน 2 คนรวมกัน
5. สถานการณ์ที่ 5 (S5): แบบจำลองทำนายผลการตรวจสอบ หากผลการทำนายอยู่ในช่วงขีดแบ่งที่ไม่มั่นใจ ให้นำความเห็นจากผู้ปฏิบัติงาน 1 คน เข้ามาตรวจสอบข้อมูล หากผู้ปฏิบัติงานและแบบจำลองมีความเห็นพ้องตรงกัน ให้สรุปผล แต่หากไม่ตรงกัน ให้นำความเห็นจากผู้ปฏิบัติงานคนที่ 2 เข้ามาร่วมตรวจสอบข้อมูล โดยตัดสินด้วยการลงคะแนนเสียงส่วนใหญ่

ในการจำลองทุกสถานการณ์ จะค้นหาขีดแบ่งที่เหมาะสมด้วยการใช้การค้นหาแบบกริด (Grid Search) ซึ่งจะจูนและค้นหาค่าขีดแบ่งโดยลองใช้ค่าขีดแบ่งในทุกรูปแบบ จากนั้นประเมินประสิทธิภาพของแบบจำลอง โดยตั้งเกณฑ์ให้เลือกค่าขีดแบ่งที่เสียค่าใช้จ่ายในการจ้างผู้ปฏิบัติงาน

น้อยที่สุด ที่ยังสามารถทำนายผลลัพธ์ด้วยความแม่นยำ (Accuracy) ที่มากกว่า 90% เมื่อได้ค่าขีดแบ่งดังกล่าวแล้ว จึงใช้แบบจำลองและค่าขีดแบ่งที่หาได้ ทำนายผลการจำแนกข้อมูลและดำเนินการทดลองตามสถานการณ์ทั้ง 5 เพื่อนำมาวิเคราะห์และพิจารณาถึงผลกระทบจากการนำมาประยุกต์ใช้จริงด้วยสถานการณ์ต่าง ๆ

ผลลัพธ์

จากผลการทดลองใช้แบบจำลองช่วยตรวจสอบคุณภาพของข้อมูลในสถานการณ์ต่าง ๆ พบว่า แบบจำลองสามารถช่วยคัดกรองข้อมูลที่มีคุณภาพต่ำออกไปได้บางส่วน และสามารถช่วยลดทรัพยากรที่ต้องเสียไปในการตรวจสอบคุณภาพของข้อมูลลงได้

ตารางที่ 13 แสดงผลการใช้แบบจำลองเอสวีเอ็มในงานอัดเสียงพูดประโยคภาษาไทยเมื่อนำมาประยุกต์ใช้ในสถานการณ์ที่แตกต่างกัน พบว่า การประยุกต์ใช้ในกรณีที่ 3, 4 และ 5 ซึ่งนำแบบจำลองเข้ามาช่วยสามารถประหยัดค่าใช้จ่ายลงไปได้สูงสุดถึง 30% โดยคุณภาพของข้อมูลสุดท้ายลดลงไปเพียงประมาณ 6% เมื่อเทียบกับสถานการณ์ที่ 2 และลดค่าใช้จ่ายลงไปได้มากถึง 89.86% เมื่อเทียบกับสถานการณ์ที่ 1 ในขณะที่ความความถูกต้องของข้อมูลสุดท้ายลดลงเพียงประมาณ 10% เท่านั้น นอกจากนี้ตารางที่ 14 เมื่อนำแบบจำลองเอ็กจีบูสต์มาใช้ในงานตรวจสอบคุณภาพคลิปเสียงพูด ก็สามารถลดค่าใช้จ่ายลงไปได้สูงสุดถึง 66.67% ในขณะที่ความแม่นยำลดลงมากที่สุดเพียง 6.69% เท่านั้น เมื่อเทียบสถานการณ์ที่ 5 กับสถานการณ์ที่ 1

	S1	S2	S3	S4	S5
จำนวนคลิป	33,577.00	33,577.00	33,577	33,577.00	33,577.00
tp	23,516.00	9,637.00	7,084.67	7,669.00	7,669.00
fp	0.00	917.33	327.33	959.00	959.00
tn	10,061.00	22,598.67	23,188.67	22,557.00	22,557.00
fn	0.00	424.00	2,976.33	2,392.00	2,392.00
precision	NaN	0.9131	0.9558	0.8889	0.8889
recall	0.0000	0.9579	0.7042	0.7623	0.7623
F1 score	NaN	0.9349	0.8109	0.8207	0.8207
Accuracy	1.0000	0.9601	0.9016	0.9002	0.9002
Ground Truth จำแนกผ่าน	23,516.00	23,516.00	23,516.00	23,516.00	23,516.00

Ground Truth จำแนกไม่ผ่าน	10,061.00	10,061.00	10,061.00	10,061.00	10,061.00
แบบจำลอง มั่นใจ จำแนกผ่าน	-	-	21,371.00	18,822.00	18,822.00
แบบจำลอง มั่นใจ จำแนกไม่ผ่าน	-	-	1,996.00	4,661.00	4,661.00
แบบจำลอง ไม่มั่นใจ จำแนกผ่าน	-	-	2,757.00	5,306.00	5,306.00
แบบจำลอง ไม่มั่นใจ จำแนกไม่ผ่าน	-	-	7,453.00	4,788.00	4,788.00
จำนวนคลิปที่ต้องใช้ผู้ปฏิบัติงาน	33,577.00	33,577.00	10,210.00	10,094.00	10,094.00
จำนวนผู้ปฏิบัติงานที่ใช้รวมเฉลี่ย	100,731.00	33,577.00	10,210.00	20,188.00	14,226.00
จำนวนค่าใช้จ่ายที่ใช้รวมเฉลี่ย	100,731.00	33,577.00	10,210.00	20,188.00	14,226.00
ค่าขีดแบ่ง จุดต่ำกว่านี้ไม่ผ่านแน่นอน	-	-	0.30	0.22	0.22
ค่าขีดแบ่ง จุดสูงกว่านี้ผ่านแน่นอน	-	-	0.94	0.78	0.78
แบบจำลอง ไม่มั่นใจ จำแนกผ่าน + ผู้ปฏิบัติงานตามสถานการณ์ แล้วผลเป็น ผ่าน	-	-	4601.67	5934.00	5934.00
แบบจำลอง ไม่มั่นใจ จำแนกผ่าน + ผู้ปฏิบัติงานตามสถานการณ์ แล้วผลเป็น ไม่ผ่าน	-	-	192.33	193.00	193.00

แบบจำลอง ไม่มั่นใจ จำแนกไม่ผ่าน + ผู้ปฏิบัติงานตามสถานการณ์ แล้วผลเป็น ผ่าน	-	-	196.33	205.00	205.00
แบบจำลอง ไม่มั่นใจ จำแนกไม่ผ่าน + ผู้ปฏิบัติงานตามสถานการณ์ แล้วผลเป็น ไม่ผ่าน	-	-	5219.67	3762.00	3762.00

ตารางที่ 13 ตารางแสดงค่าการวัดผลต่าง ๆ เมื่อจำลองสถานการณ์ทั้ง 5 โดยใช้แบบจำลองเอสวีเอ็ม ในงานที่ 1 งานอัดเสียงพูดประโยคภาษาไทย

	S1	S2	S3	S4	S5
จำนวนคลิป	20,265.00	20,265.00	20,265.00	20,265.00	20,265.00
tp	6,565.00	6,399.00	6,308.00	6,418.00	6,394.00
fp	0.00	576.00	1,098.00	212.00	412.00
tn	13,700.00	13,124.00	12,602.00	13,488.00	13,288.00
fn	0.00	166.00	257.00	147.00	171.00
precision	NaN	0.9174	0.8517	0.9680	0.8517
recall	0.0000	0.9747	0.9609	0.9776	0.9609
F1 score	NaN	0.9452	0.9030	0.9728	0.9030
Accuracy	1.0000	0.9634	0.9331	0.9823	0.9331
Ground Truth จำแนกผ่าน	13,700.00	13700.00	13,700.00	13,700.00	13700.00
Ground Truth จำแนกไม่ผ่าน	6,565.00	6,565.00	6,565.00	6,565.00	6565.00
แบบจำลอง มั่นใจ จำแนกผ่าน	-	-	609.00	13,635.00	12859.00

แบบจำลอง มั่นใจ จำแนกไม่ผ่าน	-	-	1,040.00	6,630.00	7406.00
แบบจำลอง ไม่มั่นใจ จำแนกผ่าน	-	-	12,250.00	0.00	0.00
แบบจำลอง ไม่มั่นใจ จำแนกไม่ผ่าน	-	-	6,366.00	0.00	0.00
จำนวนคลิปที่ต้องใช้ผู้ปฏิบัติงาน	20,265.00	20,265.00	18,616.00	0.00	0.00
จำนวนผู้ปฏิบัติงานที่ใช้รวมเฉลี่ย	60,795.00	20,265.00	18,616.00	20,265.00	20,265.00
จำนวนค่าใช้จ่ายที่ใช้รวมเฉลี่ย	60,795.00	20,265.00	38,880.00	20,265.00	20,265.00
ค่าขีดแบ่ง จุดต่ำกว่านี้ไม่ผ่านแน่นอน	-	-	0.00	0.32	0.06
ค่าขีดแบ่ง จุดสูงกว่านี้ผ่านแน่นอน	-	-	0.06	0.34	0.06
แบบจำลอง ไม่มั่นใจ จำแนกผ่าน + ผู้ปฏิบัติงานตามสถานการณ์ แล้วผลเป็น ผ่าน	-	-	12,158.00	0.00	0.00
แบบจำลอง ไม่มั่นใจ จำแนกผ่าน + ผู้ปฏิบัติงานตามสถานการณ์ แล้วผลเป็น ไม่ผ่าน	-	-	92.00	0.00	0.00

แบบจำลอง ไม่มั่นใจ จำแนกไม่ผ่าน + ผู้ปฏิบัติงานตามสถานการณ์ แล้วผลเป็น ผ่าน	-	-	132.00	0.00	0.00
แบบจำลอง ไม่มั่นใจ จำแนกไม่ผ่าน + ผู้ปฏิบัติงานตามสถานการณ์ แล้วผลเป็น ไม่ผ่าน	-	-	6,234.00	0.00	0.00

ตารางที่ 14 ตารางแสดงค่าการวัดผลต่าง ๆ เมื่อจำลองสถานการณ์ทั้ง 5 โดยใช้แบบจำลองเอ็กซ์จีบูสต์ในงานที่ 2 งานตรวจสอบคุณภาพของคลิปเสียงพูด

4.3.4. การวัดผลแบบจำลอง

การวัดผลแบบจำลองจะทำการคำนวณคะแนนรูปแบบต่าง ๆ ทั้ง ค่าความแม่น (Accuracy) ค่าความเที่ยง (Precision) ค่าความระลึกได้ (Recall) และค่าคะแนนเอฟวัน (F_1 score) เพื่อวัดผลแบบจำลองด้วยมุมมองที่แตกต่างกัน แต่จะให้น้ำหนักกับค่าความระลึกได้และค่าคะแนนเอฟวันเป็นพิเศษ เนื่องจากระบบตรวจหาคำตอบคุณภาพต่ำในคราวด์ซอร์สซิงแพลตฟอร์ม ควรคำนึงถึงความเสียหายที่จะเกิดขึ้นหากมีข้อมูลคุณภาพต่ำ ที่แบบจำลองดันจำแนกว่าไม่ได้มีข้อมูลคุณภาพต่ำ ซึ่งเป็นเหตุการณ์แบบ เท็จลบ (False Negative) ส่งผลให้เกิดความเสียหายขึ้น โดยข้อมูลที่ผู้ที่ต้องการข้อมูลได้รับไปจะมีคุณภาพต่ำ แต่ในขณะที่มองค่าความระลึกได้เป็นหลัก ก็ไม่ควรเน้นการตรวจหาคำตอบคุณภาพต่ำ จนแบบจำลองเอนเอียงไปทางที่จำแนกว่าพฤติกรรมนั้นจะส่งผลให้คำตอบคุณภาพต่ำมากเกินไป เพราะหากเป็นแบบนั้น จะทำให้เกิดเหตุการณ์ เท็จบวก (False Positive) มากขึ้น กล่าวคือ ข้อมูลผลลัพธ์คำตอบที่คุณภาพสูง ถูกแบบจำลองจำแนกว่ามีคุณภาพต่ำไปด้วย ซึ่งหากเกิดน้อยครั้ง ผู้ปฏิบัติงานที่ทำงานคุณภาพสูงอาจจะมีการร้องเรียนให้ตรวจสอบให้ เพื่อแก้ไขให้คำตอบถูกจำแนกกลับไปว่าคุณภาพสูงเหมือนเดิม แต่หากเกิดเหตุการณ์นี้บ่อยครั้ง จะส่งผลเสียให้ผู้ปฏิบัติงานรู้สึกไม่พอใจ และอาจเลิกปฏิบัติงานให้กับแพลตฟอร์ม ดังนั้นเราจึงควรสนใจการเกิดเหตุการณ์เท็จบวกด้วยอย่างสมดุล จึงควรพิจารณาค่าคะแนนเอฟวันประกอบด้วย

บทที่ 5 สรุปผลการวิจัย และ ข้อเสนอแนะ

5.1. อภิปรายและสรุปผล

การควบคุมคุณภาพข้อมูลของคราวด์ซอร์สซิงแพลตฟอร์มเป็นสิ่งสำคัญ ในงานนี้ได้วิเคราะห์ประสิทธิภาพของวิธีควบคุมคุณภาพด้วยการเพิ่มกระบวนการในการเก็บข้อมูลที่แตกต่างกัน ได้แก่ 1. งานที่จำเป็นต้องทำก่อน 2. คำถามมาตรฐานทองคำ และ 3. การทำซ้ำของข้อมูล

งานที่จำเป็นต้องทำก่อน (Job Prerequisite) ใช้คัดกรองผู้ปฏิบัติงานที่มีความรู้ ทักษะ ความสามารถ ความเข้าใจและคุณสมบัติที่เหมาะสมและเป็นไปตามต้องการสำหรับงานนั้น งานวิจัยนี้ได้วิเคราะห์คำถามทั้งหมด 3 ประเภท ประกอบด้วย 1. คำถามที่สามัญสำนึก (Common Sense Questions) 2. คำถามเกี่ยวกับการดำเนินงาน (Logistical Questions) และ 3. คำถามเกี่ยวกับรายละเอียดงาน (Job Detail Questions) คำถามเกี่ยวกับรายละเอียดงานเป็นประเภทที่มีประสิทธิภาพมากที่สุดในการคัดกรองผู้ปฏิบัติงาน

คำถามมาตรฐานทองคำ (Gold Standard Question) วิธีนี้เป็นกระบวนการแทรกคำถามวัดผลเข้าไปในงานระหว่างกระบวนการทำงานเพื่อคัดกรองข้อมูลที่มีคุณภาพต่ำออกไป งานวิจัยนี้ได้วิเคราะห์คำถาม 2 ประเภท คือ คำถามที่ชัดเจน (Obvious Questions) และคำถามที่ซ้ำกัน (Duplicate Questions) ประเภทของคำถามที่ซ้ำกันมีประสิทธิภาพมากกว่าในการคัดกรองข้อมูลที่มีคุณภาพต่ำ

การทำซ้ำของข้อมูล เป็นกระบวนการฉันทามติซึ่งเป็นประโยชน์ในการเพิ่มคุณภาพของคำตอบ จากผลการทดลองพบว่าหากข้อมูลถูกเก็บรวมจากผู้ปฏิบัติงานที่ผ่านการคัดกรองแล้ว ไม่จำเป็นต้องใช้การทำซ้ำของข้อมูล โดยผู้ปฏิบัติงานคนเดียวก็เพียงพอที่จะได้ข้อมูลที่มีคุณภาพสูงถึง 94.06% อย่างไรก็ตาม ในบางกรณี เช่น งานทางการแพทย์ซึ่งต้องการความแม่นยำสูง อาจจำเป็นต้องใช้ทรัพยากรเพิ่มเติมในการเก็บรวบรวมข้อมูลด้วยการทำซ้ำของข้อมูลเพื่อให้ได้ข้อมูลที่มีคุณภาพสูงขึ้น

แต่ละวิธีการถูกดำเนินการในขั้นตอนที่แตกต่างกันของกระบวนการเก็บรวบรวมข้อมูลและเหมาะสมสำหรับสถานการณ์ที่แตกต่างกัน ดังนั้น จึงสามารถนำมาประยุกต์ใช้พร้อมกันในงานเดียวได้ อย่างไรก็ตาม การเพิ่มขั้นตอนการควบคุมคุณภาพหรือเพิ่มจำนวนคำถามในแต่ละขั้นตอนมักจะเพิ่มประสิทธิภาพในการคัดกรองข้อมูลที่มีคุณภาพต่ำหรือผู้ปฏิบัติงานที่มีคุณภาพต่ำ แต่ก็ต้องแลกมาด้วยค่าใช้จ่ายเพิ่มเติมและเวลาในการเก็บรวบรวมข้อมูลที่เพิ่มขึ้น การแลกเปลี่ยน (Trade-off) นี้ควรได้รับการคำนึงถึงอย่างเหมาะสมโดยพิจารณาผลลัพธ์ที่ต้องการและทรัพยากรที่มีอยู่เป็นปัจจัยสำคัญ ปัจจัยที่ควรคำนึงถึง เช่น คุณภาพของข้อมูลที่ต้องการ งบประมาณที่มี และระยะเวลาการส่งมอบข้อมูลที่ต้องการ การคำนึงถึงปัจจัยเหล่านี้สามารถช่วยผู้ร้องขอข้อมูลให้กำหนดขอบเขตของการออกแบบการเก็บรวบรวมข้อมูลและเลือกใช้วิธีการควบคุมคุณภาพที่เหมาะสม

นอกจากนี้ การใช้แบบจำลองการเรียนรู้ของเครื่องในการตรวจสอบพฤติกรรมการทำงานของผู้ปฏิบัติงานเพื่อทำนายและจำแนกคุณภาพของข้อมูลที่ได้รับ ยังสามารถช่วยเพิ่มประสิทธิภาพการคัดกรองข้อมูลคุณภาพต่ำออกไปได้ ในขณะที่ยังช่วยลดปริมาณทรัพยากรที่ต้องสูญเสียในการเก็บข้อมูลและตรวจสอบคุณภาพของข้อมูลลงอีกด้วย ดังนั้นจึงเป็นการดีหากสามารถใช้แบบจำลองการเรียนรู้ของเครื่องเข้ามาตรวจสอบคุณภาพของข้อมูลที่เก็บด้วยคราวด์ซอร์สซิงแพลตฟอร์ม

ผลการศึกษาของวิทยานิพนธ์นี้สามารถใช้เป็นปัจจัยในการพิจารณาออกแบบกระบวนการเก็บรวบรวมข้อมูลในสถานการณ์ที่เก็บข้อมูลผ่านคราวด์ซอร์สซิงแพลตฟอร์ม โดยเป็นแนวทางพื้นฐานสำหรับผู้ร้องขอข้อมูลเพื่อเริ่มวางแผนการออกแบบงานของตนเอง

5.2. ข้อเสนอแนะสำหรับงานวิจัยในอนาคต

สำหรับงานวิจัยในอนาคต สิ่งสำคัญประการหนึ่งคือ ต้องตรวจสอบผลกระทบของจำนวนงานต่อผู้ปฏิบัติงานและผลกระทบต่อความรู้สึกของผู้ปฏิบัติงานนั้น นอกจากนี้ ยังควรคำนึงถึงค่าตอบแทนเทียบกับปริมาณงานและจำนวนผู้ปฏิบัติงานด้วย วิธีการตรวจสอบและควบคุมคุณภาพของข้อมูลอื่น ๆ ยังสามารถศึกษาและนำมาใช้ในงานได้ เช่น การแบ่งงานออกเป็นหลายส่วนย่อยโดยใช้ปัจจัยความซ้ำซ้อนที่แตกต่างกัน การตรวจสอบที่ใช้การเรียนรู้ของเครื่องด้วยแบบจำลองอื่น ๆ การประยุกต์ใช้การเพิ่มกระบวนการและการใช้แบบจำลองการเรียนรู้ของเครื่องเข้าด้วยกันในงานเดียว สิ่งเหล่านี้จะให้ข้อมูลเพิ่มเติมเกี่ยวกับการแลกเปลี่ยนที่เป็นไปได้ และความสามารถของการตัดสินใจสำหรับการออกแบบการเก็บรวบรวมข้อมูลสำหรับงานที่แตกต่างกันในอนาคต

บรรณานุกรม

1. Mourelatos, E., M. Tzagarakis, and E.J.S.-E.E.J.o.E. Dimara, *A review of online crowdsourcing platforms.* 2016. **14**(1).
2. Seber, G.A. and A.J. Lee, *Linear regression analysis*. Vol. 329. 2012: John Wiley & Sons.
3. Mok, R.K., W. Li, and R.K. Chang. *Detecting low-quality crowdtesting workers*. in *Quality of Service (IWQoS), 2015 IEEE 23rd International Symposium on*. 2015. IEEE.
4. Hirth, M., T. Hoßfeld, and P. Tran-Gia. *Cost-optimal validation mechanisms and cheat-detection for crowdsourcing platforms*. in *Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2011 Fifth International Conference on*. 2011. IEEE.
5. Buchholz, S. and J. Latorre. *Crowdsourcing preference tests, and how to detect cheating*. in *Twelfth Annual Conference of the International Speech Communication Association*. 2011.
6. Hirth, M., et al. *Predicting result quality in crowdsourcing using application layer monitoring*. in *Communications and Electronics (ICCE), 2014 IEEE Fifth International Conference on*. 2014. IEEE.
7. Chen, M.C., J.R. Anderson, and M.H. Sohn. *What can a mouse cursor tell us more?: correlation of eye/mouse movements on web browsing*. in *CHI'01 extended abstracts on Human factors in computing systems*. 2001. ACM.
8. Mok, R.K., W. Li, and R.K.J.A.S.C.C.R. Chang, *A user behavior based cheat detection mechanism for crowdtesting.* 2015. **44**(4): p. 123-124.
9. Mok, R.K., R.K. Chang, and W.J.I.T.o.M. Li, *Detecting low-quality workers in QoE Crowdtesting: a worker behavior-based approach.* 2017. **19**(3): p. 530-543.
10. Kazai, G. and I. Zitouni. *Quality management in crowdsourcing using gold judges behavior*. in *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. 2016. ACM.
11. Matsuda, Y., Y. Suzuki, and S. Nakamura. *A trade-off between estimation*

accuracy of worker quality and task complexity. in *Big Data (Big Data), 2017 IEEE International Conference on*. 2017. IEEE.

12. Lyu, L., *Spam elimination and bias correction: ensuring label quality in crowdsourced tasks.* 2018.
13. Schroff, F., D. Kalenichenko, and J. Philbin. *Facenet: A unified embedding for face recognition and clustering*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.
14. Bromley, J., et al. *Signature verification using a" siamese" time delay neural network*. in *Advances in neural information processing systems*. 1994.

ประวัติผู้เขียน

ชื่อ-สกุล	กฤตย์ กังวาลพงศ์พันธุ์
วัน เดือน ปี เกิด	23 พฤศจิกายน 2537
สถานที่เกิด	กรุงเทพมหานคร
วุฒิการศึกษา	จุฬาลงกรณ์มหาวิทยาลัย
ที่อยู่ปัจจุบัน	43/27 หมู่บ้านพระราม 4 ซอยพระยาพิเรนทร์ ถนนพระราม 4 – สุวรรณสวัสดิ์ แขวงทุ่มมหาเมฆ เขตสาทร กรุงเทพฯ 10120
ผลงานตีพิมพ์	Teaching Methods in Engineering Education: A Case Study in Thailand Analysis of the Trade-Off in Data Quality Management for Crowdsourcing